Proceedings of the 6th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-6)

Bunt, H.C.

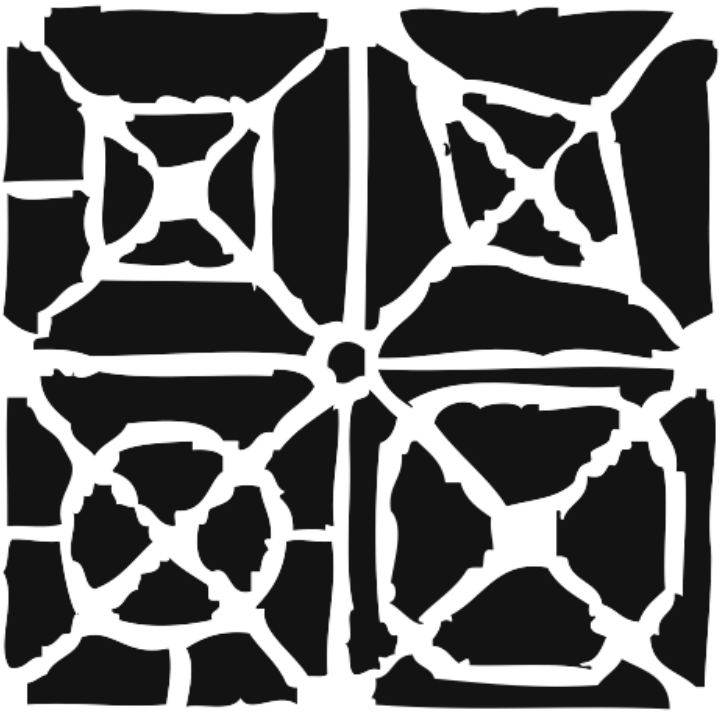
Publication date:
2011

Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation

isa-6

January 11–12, 2011

Oxford, UK



TiCC, Tilburg Center for Cognition and Communication

Tilburg Center for Cognition and Communication
and Communication
Tilburg University
http://www.uvt.nl/ticc

P.O. Box 90153
5000 LE Tilburg
The Netherlands
ticc@uvt.nl

January 11, 2011

TiCC TR 2011-002

Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation

isa-6

Computing Laboratory, University of Oxford, UK

In conjunction with the 9th International Conference

on Computational Semantics IWCS 2011

Harry Bunt, editor

TiCC, Tilburg Center for Cognition and Communication
ISBN: 978-90-74029-35-3

Contents

Preface

This slender volume contains the accepted long and short papers that were submitted to the Sixth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, isa-6, which was organized in Oxford, UK, January 11-12, 2011 in conjunction with is the Nineth International Worksop on Computational Semantics (IWCS 2011).

isa-6 is the sixth edition of joint workshops on the International Organization for Standards ISO and the ACL Special Interest Group in Computtional Semantics, Working Group "The Representation of Multimodal Semantic Information (`http://sigsem.uvt.nl`). The isa workshops are organized on the occasion of meetings of ISO projects concerned with the establishment of international standards for semantic annotation and representation. The main focus of these workshops is on the presentation and discussion of approaches, experiments, experiences, and proposals concerning the construction or application of interoperable linguistic resources with semantic annotations.

The isa-6 workshop co-occurs with meetings of several subprojects of the ISO project 24617, "Semantic annotation framework (SemAF)", concerned with dialogue act annotation, the annotation of spatial information, the annotation of semantic roles, the annotation of discourse relations, and the annotation of named entities, as well as the relation to ontologies.

I would like to thank the members of the isa-6 Programme Committee for their careful and quick reviewing; the IWCS 2011 organizers for their support, and the members of the isa-6 organizing committee for smooth cooperation.

Harry Bunt
isa-6 chair

Programme committee members

Jan Alexandersson (DFKI, Saarbrücken)
Antal van den Bosch (Tilburg University)
Harry Bunt (Tilburg University)
Nicoletta Cakzolari (ILC-CNR, Pisa)
Jae-Woong Choe (Korea Uiversity, Seoul)
Thierry Declerck (DFKI, Saarbrücken)
Alex Chengyu Fang (City University of Hong Kong)
Gil Francopoulo (Tagmatica, Paris)
Koiti Hasida (AIST, Tokyo)
Nancy Ide (Vassar College, Poughkeepsie, NY)
Kiyong Lee (Korea University, Seoul)
Lluis Marquez (UPC, Barcelona)
Martha Palmer (University of Colorado, Boulder, Co.)
Volha Petukhova (Tilburg University)
Massimo Poesio (University of Trento)
Andrei Popescu-Belis (IDIAP, Martigny, Switzerland)
James Pustejovsky (Brandeis University, Waldham, MA)
Laurent Romary (CNRS, Berlin)
David Traum (USC ICT, Marina del Rey, Cal.)
Piek Vossen (VU Amsterdam)
Yorick Wilks (University of Sheffield)

Organizers

Harry Bunt (Tilburg University)
Kiyong Lee (Korea University, Seoul)
Gil Francopoulo (Tagmatica, Paris)
Stephen Pulman (University of Oxford)
Martha Palmer (University of Colorado, Boulder, CO)
James Pustejofsky (Brandeis University, Waldham, MA)
Laurent Romary (CNRS, Berlin)

Workshop program

Venue: Computing Laboratory, University of Oxford, UK

TUESDAY January 11, 2011

09.00 - 09.10	Opening
09.10 - 09.40	James Pustejovsky, Jessica Moszkowics and Marc Verhagen: *ISO-Space: The Annotation of Spatial Information in Language*
09.40 - 10.45	Discussion on ISO PWI 24617-6 ISO-Space (1) (PL: James Pustejovsky)
10.45 - 11.00	coffee break
11.00 - 12.10	Discussion on ISO PWI 24617-6 ISO-Space (2)
12.10 - 12.30	Leon Derczynski and Robert Gaizauskas: *An Annotation Scheme for Reichenbachs Verbal Tense Structure*
12.30 - 14.00	lunch break
14.00 - 15.15	Discussion on ISO DIS 24617-2 Dialogue acts (1) (PL: Harry Bunt)
15.15 - 15.30	tea break
15.30 - 16.15	Discussion on ISO DIS 24617-2 Dialogue acts (2)
16.15 - 16.45	Volha Petukhova, Laurent Prvot and Harry Bunt: *Multi-level Discourse Relations Between Dialogue Units*
16.45 - 17:45	Discussion on ISO PWI 24622 Lexical ontology (PL: Thierry Declerck)
17.45 - 18.00	Thierry Declerck, Piroska Lendvai and Tobias Wunner: *Linguistic and Semantic Description of Labels Used in Knowledge Representation Systems*

WEDNESDAY January 12, 2011

09.00 - 09.45	Discussion on ISO PWI 24617-2 SemAF-Named Entities (PL: Gil Francopoulo)
09.45 - 10.00	Gil Francopoulo and Franois Demay: *A Deep Ontology for Named Entities*
10.00 - 10.20	Claire Bonial, Susan Windisch Brown, William Convey, Martha Palmer, Volha Petukhova and Harry Bunt: *An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS*
10.20 - 10.35	Abdelati Hawwar, Jena D. Hwang, Aous Mansouri and Martha Palmer: *Classification and Deterministic PropBank Annotation of Predicative Adjectives in Arabic*
10.35 - 10.50	coffee break
10.50 - 12.30	Discussion on ISO WD 24617-4 SemAF-Semantic Roles (PL: Martha Palmer)
12.30 - 14.00	lunch break
14.00 - 14.20	Nancy Ide, Rashmi Prasad and Aravind Joshi: *Towards Interoperability for the Penn Discourse Treebank*
14.20 - 15.20	Discussion on ISO WD 24617-5 SemAF-Discourse Structures (1) (PL: Koiti Hasida)
15.20 - 15.35	tea break
15.35 - 16.15	Discussion on ISO WD 24617-5 SemAF-Discourse Structures (2)
16.15 - 16.45	Discussion on ISO PWI 24617-X SemAF-Basics (PL: Harry Bunt)
16.45 - 16.50	Closing of ISA-6

ISO-Space: The Annotation of Spatial Information in Language

James Pustejovsky
Brandeis University
jamesp@cs.brandeis.edu

Jessica L. Moszkowicz
Brandeis University
jlittman@cs.brandeis.edu

Marc Verhagen
Brandeis University
marc@cs.brandeis.edu

Abstract

We introduce ISO-Space, an annotation specification for capturing spatial and spatiotemporal information in natural language. We discuss many of the issues found in spatial language and show how ISO-Space aims to address these problems. ISO-Space is an emerging resource that is still in its early stages of development. We describe the genres of text that will be used in a pilot annotation study, in order to refine and enrich the specification language.

1 Motivation and Problem Definition

Natural languages are filled with particular constructions for talking about spatial information, including spatially anchored events, locations that are described in relation to other locations, and movement along a path. While representing and reasoning about spatial information has recently received ample attention, particularly from the qualitative reasoning community, that work often overlooks the complexity that language brings to the problem. In fact, establishing tighter formal specifications of the relationship between language and space has proved to be a considerable challenge. In this paper, we propose an annotation framework called ISO-Space that aims to be such a specification. ISO-Space incorporates the annotations of static spatial information, borrowing from the SpatialML scheme (MITRE, 2007; Mani et al., 2008), along with a new markup language called Spatiotemporal Markup Language (STML) (Pustejovsky and Moszkowicz, 2008) that focuses on locating events in space.

The name "ISO-Space" is used in particular because this markup language is being developed within the ISO TC37/SC4 technical subcommittee on language resource management as part six of the Semantic Annotation Framework, where the goal is to create a new standard for capturing spatial and spatiotemporal information.

In previous analyses of spatial information, it has been assumed that language makes use of a relatively simple inventory of terms in order to describe spatial information. In approaches such as these, the principal burden of explanation is located within non-linguistic formalisms, but such characterizations are ill-suited for dealing with the extreme flexibility of spatial language as used in real contexts.

There are many applications and tasks which would benefit from a robust spatial markup language such as ISO-Space. These include:

- Building a spatial map of objects relative to one another.
- Reconstructing spatial information associated with a sequence of events.
- Determining object location given a verbal description.
- Translating viewer-centric verbal descriptions into other relative descriptions or absolute coordinate descriptions.
- Constructing a route given a route description.
- Constructing a spatial model of an interior or exterior space given a verbal description.
- Integrating spatial descriptions with information from other media.

To this end, the goal of ISO-Space is not to provide a formalism that fully represents the complexity of spatial language, but rather to capture these complex constructions in text to provide an inventory of how spatial information is presented in natural language. The framework is built on previous and ongoing work on annotations for spatial, temporal, and spatiotemporal information, but we also follow the MATTER cycle as described in

(Pustejovsky, 2006) and illustrated below in Figure 1. Following that strategy, we aim to look frequently at real text and adjust the specification of ISO-Space accordingly after several rounds of annotation.

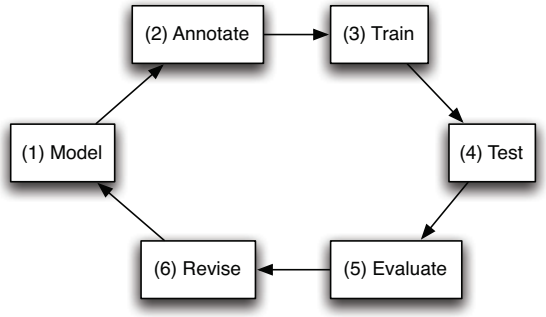


Figure 1: MATTER Development Cycle

In this paper, we first describe the semantic requirements for the annotation language, and then discuss the structure of the annotation framework. We then outline the basic elements of the current version of ISO-Space, followed by explicit examples of how these elements are used for markup. We briefly discuss our strategy of corpus-driven development of the specification, and conclude with remaining issues and outstanding questions of interpretation.

2 Semantic Requirements for Annotation

We follow the ISO CD 24612 *Language resource management - Linguistic Annotation Framework* proposed standard, in making a fundamental distinction between the concepts of *annotation* and *representation* (Ide and Romary, 2004). This distinction is reflected in the current development of ISO-Space, where we are taking great care to distinguish between an *abstract syntax* and a *concrete syntax*. While the concrete syntax is exemplified by a specific XML encoding, for example, the abstract syntax defines the model and the structures constituting the information about space, directions, and movement that may be contained in annotations. The abstract syntax consists of a *conceptual inventory* (Bunt and Pustejovsky, 2010) and a set of syntactic rules defining the combinations of these elements. While still in development, it is clear that the conceptual inventory for spatial language annotation must include (at least) the following notions:

- Locations (regions, spatial objects): Geographic, Geopolitical Places, Functional Locations.
- Entities viewed as Spatial Objects.
- Paths as Objects: routes, lines, turns, arcs
- Topological relations: *inside*, *connected*, *disconnected*.
- Direction and Orientation: *North*, *downstream*.
- Time and space measurements: units and quantities for spatial and temporal concepts.
- Object properties: intrinsic orientation, dimensionality, size, shape.
- Frames of reference: absolute, intrinsic, relative.
- Spatial Functions: *behind the building*, *twenty miles from Boulder*.
- Motion: tracking moving objects over time.

It is these concepts which are specified set-theoretically in the abstract syntax, and for which a formal semantics must be provided (Bunt and Romary, 2002). In the present paper, however, we focus mainly on general characteristics and requirements for an annotation language of spatial information and hence, we will have little to say regarding the semantics.

We will refer to constructions that make explicit reference to the spatial attributes of an object or spatial relations between objects as *spatial expressions*. Linguists traditionally divide spatial expressions into at least four grammatically defined classes:

(1) a. Spatial Prepositions and Particles: *on*, *in*, *under*, *over*, *up*, *down*, *left of*;

b. Verbs of position and movement: *lean over*, *sit*, *run*, *swim*, *arrive*;

c. Spatial attributes: *tall*, *long*, *wide*, *deep*;

d. Spatial Nominals: *area*, *room*, *center*, *corner*, *front*, *hallway*.

Unlike the fairly well-behaved list of 13 values for temporal relations in language (as encoded in ISO-TimeML), spatial prepositions are notoriously ambiguous and context dependent. Not only are there vastly more configurations possible between objects construed as spatial regions, but languages are idiosyncratic in how spatial information is encoded through different linguistic expressions. For

this reason, we will have to define constraints that allow for underspecified semantic interpretations for several of the concepts introduced in our abstract syntax. These will need to communicate with various lexical (Miller, 1995; Fellbaum, 1998; Kipper et al., 2006) and spatial ontological resources (Frank, 1997; Bateman et al., 2010), in order to help disambiguate and more fully determine the semantics of relation types from the specification (Grenon and Smith, 2004).

3 The Annotation Framework

ISO-Space is designed to capture two kinds of information: spatial information and spatiotemporal information. To accomplish this, we bring together three existing resources, as shown below in Figure 2.

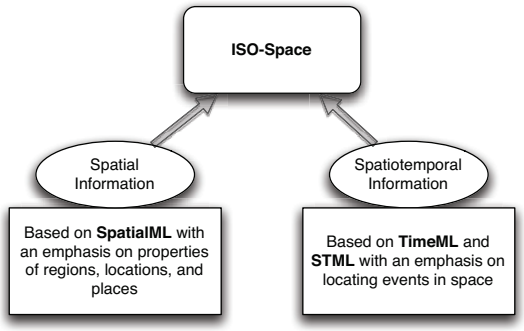


Figure 2: ISO-Space Components

There has been considerable research on the linguistic behavior of spatial predicates and prepositions in language (Talmy, 1983; Jackendoff, 1983; Herskovits, 1986; Boas, 2001; Cappelle and Declerck, 2005). Within qualitative spatial reasoning (QSR), work has recently started to focus on incorporating mereo-topological concepts into the calculus of relations between regions (Randell et al., 1992; Herring et al., 1994; Egenhofer et al., 1994; Kurata and Egenhofer, 2007).

The focus of SpatialML is to mark up spatial locations mentioned in texts while allowing integration with resources that provide information about a given domain, such as physical feature databases and gazetteers. The core SpatialML tag is the PLACE tag, which includes attributes `type` (country, continent, populated place, building, etc.), `gazref` (a reference to a gazetteer entry), and `LatLong` (a latitude and longitude pair). Complex locations such as *Pacific coast of Australia* and *the hot dog stand behind Macy's* are annotated using the LINK and RLINK tags, respectively, encoding topological and relative relations between places.

SpatialML is one of the cornerstones of ISO-Space, but it needs to be extended to account for more of the complexity of spatial language. Most notably, SpatialML was not designed to capture implicit places such as the one described by the phrase *behind Macy's* rather than the more complete example *the hot dog stand behind Macy's*.

In addition to capturing implicit spatial information, ISO-Space will include additional properties of locations such as orientation and metric relations between objects, the shape of an object, the size of an object, elevation, granularity, aggregates and distributed objects, and objects in motion.

One of the most important aspects of the new framework is to deal with the full range of verbal descriptions of spatial relations. Spatial relations as expressed in language employ a combination of five parameters of meaning:

- topological: *in*, *inside*, *touching*, *outside*;
- orientational (with frame of reference): *behind*, *left of*, *in front of*;
- topo-metric: *near*, *close by*;
- topological-orientational: *on*, *over*, *below*;
- metric: *20 miles away*

The other components of ISO-Space, STML and TimeML/ISO-TimeML (Pustejovsky et al., 2005; Pustejovsky et al., 2010), bring in the spatiotemporal dimension. This dimension is captured by introducing spatial events, most specifically by annotating motions. The Spatio-Temporal Markup Language proposes a set of motion classes based on the classifications by (Muller, 1998), see Table 1.

Move	run, fly, drive
Move_External	drive around, pass
Move_Internal	walk around the room
Leave	leave, desert
Reach	arrive, enter, reach
Detach	take off, pull away
Hit	land, hit
Follow	follow, chase
Deviate	flee, run from
Stay	remain, stay

Table 1: The ten motion classes from STML

The meanings for each of these classes will correspond to a semantic interpretation of motion

concepts specified in the abstract syntax, as mentioned above.

4 The Elements of ISO-Space

There are two major types of elements distinguished in ISO-Space:

1. ENTITIES: location, spatial_entity, motion, event, path;
2. SPATIAL RELATIONS: topological, orientational, and distance.

Note that in other annotation levels, events and motions may be considered relations. In ISO-Space though, they are considered elements that can be spatially related.

Along with these two main classes, there are some minor elements, such as spatial signals, described below. In the discussion that follows, we discuss these elements in more detail, beginning with the entities. Note that all ISO-Space tags include a mandatory `id` attribute so that they can be referred to by other parts of the annotation. This tag is omitted in the discussion that follows.

Location. An *inherently grounded spatial* entity, a location includes geospatial entities such as countries, mountains, cities, rivers, etc. It also includes classificatory and ontological spatial terms, such as *edge*, *corner*, *intersection*, and so forth. The attributes for the LOCATION tag are largely inherited from SpatialML's PLACE element. For example, for those locations that have known latitude and longitude values, the `latLong` attribute must be used to allow for mapping to other resources such as Google Maps. New to ISO-Space is the *Document Creation Location* or `DCL`. This is a special location that serves as the "narrative location". If the document includes a DCL, it is generally specified at the beginning of the text, similarly to the manner in which a Document Creation Time (DCT) is specified in ISO-TimeML. If a location is the DCL, this is marked with a special attribute in the annotation of the location. The following list is a sample of some of the vetted LOCATION attributes:

- type: BODYOFWATER, CIVIL, CONTINENT, COUNTRY, LATLONG, MNT, MTS, POSTALCODE, POSTBOX, PPL, PPLA, PPLC, RGN, ROAD,
- latlong
- dcl: true, false

The values for the `type` attribute are identical to the values from the SpatialML PLACE tag, except that VEHICLE and FAC (facility) were eliminated since these are now annotated with the SPATIAL_ENTITY tag, described next.

We are pragmatic on whether there should be a distinction between place, location and region. Region, as a geometric concept, is currently not a first-class citizen in ISO-Space. The LOCATION element covers both locations and places (where a place is considered a functional category), and is assumed to be associated with a region whenever appropriate. We do acknowledge however that in some cases locations are mapped to lines instead of regions.

Spatial_entity. An entity that is not inherently a LOCATION, but one which is identified as participating in a spatial relation is tagged as a SPATIAL_ENTITY. It can be an entity such as a *car* or *building* or an individual like *John*. It can also be an event-like entity like *traffic jam* or *hurricane*. A SPATIAL_ENTITY is only annotated in the context of an explicit spatial relation. Each SPATIAL_ENTITY inherently defines a location and can be the location for other spatial entities, as in *John is in the car*. This raises the issue of whether entities like *building* in *The statue is in the building* are annotated as locations or spatial entities. We resolve this by stipulating that these entities are never annotated as locations but always as spatial entities, even in a case like *the president is in the building*.

Like LOCATION, SPATIAL_ENTITY has a `type` attribute with the following possible values: FAC, VEHICLE, PERSON, and DYNAMIC EVENT.

Motion. A MOTION is an *inherently spatial* event, involving a change of location. This element is identified in ISO-TimeML as an EVENT, but, given the spatiotemporal importance of these events, they are annotated in ISO-Space with attributes that are specific to the requirements of the present specification language; e.g., whether the event is a manner-of-motion or path predicate. Interoperability of the two specification languages should allow this element to be unified with the event attributes from ISO-TimeML, while retaining the attributes from ISO-Space as extensions.

Motions generate and participate in EVENT_PATH elements (see below). They

also have the additional attributes motion_type and motion_class, where the former allows the distinction between path and manner predicates and where the latter includes one of the STML motion classes:

- motion_type: MANNER,PATH
- motion_class: MOVE, MOVE_EXTERNAL, MOVE_INTERNAL, LEAVE, REACH, DETACH, HIT, FOLLOW, DEVIATE, STAY
- speed

Event. An EVENT is a situation that does not involve a change of location, but one which is identified as participating in a spatial relation. This element is also inherited from ISO-TimeML, with additional attributes identified for ISO-Space such as type with values STATE, PROCESS, and TRANSITION.

Note that some ISO-TimeML events are actually annotated as spatial entities in ISO-Space. For example, in *The new tropical depression was about 430 miles (690 kilometers) west of the southernmost Cape Verde Island*, the phrase *tropical depression* is annotated as a spatial entity, but ISO-TimeML has annotated it as an event.

The adoption of the ISO-TimeML EVENT element is predicated on the ability to extend the ISO-TimeML definition. For ISO-Space purposes, the definition of event may need to be enriched, but this does not necessarily have to be an ISO-TimeML responsibility.

Event_path. The implicit path that is introduced by virtue of a motion event is captured with the EVENT_PATH tag. It is defined by the motion event as well as by elements introduced in the immediate context of the motion event, like begin point, end point, path region, and intermediate points. The sourceID attribute holds the ID of the motion event. When possible the location IDs for the endpoints and a spatial signal ID, if applicable, are given in the EVENT_PATH element, which includes the following attributes:

- sourceID: motion identifier
- objectID: identifier of the moving object, typically a spatial entity
- startID: identifier of the entity that is at the beginning of the path
- endID: identifier of the entity that is at the end of the path
- end_reached: true, false
- midIDs: list of identifiers of midpoints
- path_regionID: identifier of a location
- direction: signal id
- length: distance link between the begin and end point
- signalIDs: list of signal IDs

All of these attributes, except for id and sourceID can be empty. The end_reached boolean is used to distinguish between cases like *John left for Boston* and *John arrived in Bosten*, where Boston was reached in the latter case but not in the former. The path_regionID attribute is motivated by examples like *He took I-95 towards New York*, where the event path refers to a path, but where that path is not to be equated with the event path or part of the event path.

The remaining ISO-Space tags capture information about spatial relations between the entities introduced above.

Qualitative spatial link. A *qualitative spatial link* or QSLINK is a topological or relative spatial relation between locations, spatial entities, events and motions). It includes the following attributes:

- type: TOPOLOGICAL, RELATIVE
- relation: {RCC8+}, {ORIENTATION}
- locationID: identifier of the entity that is being related
- related_locationID: identifier of entity that is being related to
- spatial_signalID: ssid

The type attribute determines what kind of relationship is annotated, distinguishing between topological and relative relations.

A *topological* QSLINK has a relation value from the extended RCC8 set which includes all RCC8 relations such as EC (touching), DC (disconnected), and PO (partially overlapping), as well as the IN relation introduced by SpatialML which is a disjunction of the RCC8 relations that imply that one region is contained within the other (TPP, NTPP, or EQ).

A *relative* QSLINK is one that essentially defines one region in terms of another and has a relation value from a closed set with position markers and direction markers. Typically, these markers are normalized versions of function words or spatial measures in the text. For

example, if the function word is *northwest of*, then the `relation` for the QSLINK will be NW. Other examples for relative relations are LEFT_OF and BEHIND. A relative QSLINK is generally introduced by a spatial signal. The attribute `spatial_signalID` holds the ID number for that ISO-Space element.

Distance. A *distance link* relates two locations, spatial entities or events and specifies the distance between them. In addition to the IDs of the relevant spatial entities, this tag asl includes `measure` and `unit` attributes.

Spatial Signal. Finally, as in ISO-TimeML, we identify the signal (or trigger) that introduces a relation with the S_SIGNAL tag. A *spatial signal* is a relation word that locates a state or motion in a particular region or path, adds information to an existing event path, or provides information on the relation type of a spatial relation.

5 Example ISO-Space Annotation

To help illustrate the ISO-Space elements introduced in the previous section, we will now look at specific natural language examples of spatial descriptions.

A simple diagram is shown in Figure 3, where the event *lives* is linked to the location *Boston* by a topological spatial link and where the spatial inclusion of the event in the location is signaled by the preposition *in*.

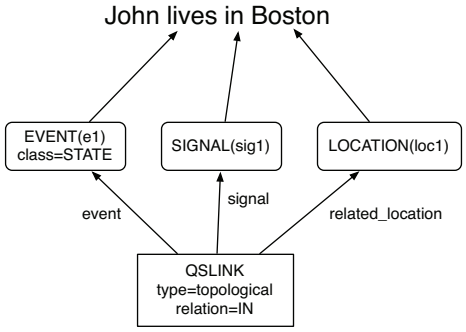


Figure 3: A simple spatial link

Two examples of sentences with spatial annotation are given in (a) and (b) in Figure 4, using inline XML to illustrate the use of tags and attributes. Note though that ISO-Space follows the Linguistic Annotation Format (Ide and Romary, 2003; Ide and Romary, 2004) in requiring stand-off annotations, the inline XML is only for purposes of exposition.

In Figure 4(a), a spatial entity is related to a location by both a topological and a relative spatial link. This may look odd but it is perfectly fine since the two relations are not contradictory. However, guidelines for annotation may stipulate that when a spatial link like `qsl1` is added, then there is no need to add `qsl2` since the DC relation may be derived from the semantics of WEST.

In 4(b), the *moving* motion creates an event path. The EVENT_PATH elements keeps a reference to the motion event in its `sourceID` attribute. In addition to the source motion, this event path specifies the moving object and the direction of the path. If possible, the location identifiers for begin and end points would have been added as well as a series of other attributes.

The text in figure 4(c) contains a variety of events, motions, spatial entities and locations. Rather than using inline XML we have marked consuming tags with square brackets and identifiers that specify the type of object. The example does not include any relations. Most annotations are relatively straightforward but we want to bring up a few points.

The noun *ship* is annotated as a spatial entity because it can be spatially related to the location *Egyptian port of El Arish*, but its modifier *Libyan* is not annotated since this particular property would not help in any spatial inference. This example brings up an interesting issue since instead of *ship* one could also link the event $docked_{e1}$ to the location. The idea is that in ISO-Space both versions would be possible and semantically equivalent given the correct argument linking and event semantics for *docked*. In general, we suggest annotating the relation between the event and the location, except when the event is a copula.

There is a grey area between events and motions. For example, is *docked* a motion that involves a movement or is it an event that is spatially situated in El Arish? This is an issue that needs to be resolved in annotation guidelines and the answer depends on the goal of the annotation. Similarly, the decision to annotate *break* as a motion rather than an event is not altogether straightforward.

The phrase *shifted course* is an interesting one. It is a motion but its exact nature is totally un-

(a) *The new tropical depression was about 430 miles (690 kilometers) west of the southernmost Cape Verde Island, forecasters said.*

The new <SPATIAL_ENTITY id="se1">tropical depression</SPATIAL_ENTITY> was about <SIGNAL id="sig1">430 miles</SIGNAL> (<SIGNAL id="sig2">690 kilometers</SIGNAL>) <SIGNAL id="sig3">west</SIGNAL> of the <LOCATION id="loc1">southernmost Cape Verde Island</LOCATION>, forecasters said.

<QSLINK id="qsl1" type="relative" entityID="s1" related_locationID="loc1" relation="WEST" signalID="sig3" />
<QSLINK id="qsl2" type="topological" entityID="s1" related_locationID="loc1" relation="DC" />
<DISTANCE id="dis1" entityID="s1" related_locationID="loc1" unit="mi" measure="430" signalID="sig1" />
<DISTANCE id="dis2" entityID="s1" related_locationID="loc1" unit="km" measure="690" signalID="sig2" />

(b) *The depression was moving westward at about 17 mph (28 kph)*

The <SPATIAL_ENTITY id="se1" type="dynamic event">depression</SPATIAL_ENTITY> was <MOTION id="m1">moving</MOTION> <SIGNAL id="sig1">westward< /SIGNAL> at about <SIGNAL id="sig2">17 mph</SIGNAL> (<SIGNAL id="sig2">28 kph</SIGNAL>)

<EVENT_PATH id="ep1" source="m1" direction="WEST" moving_object="se1" signals="[sig1]" />

(c) A Libyan [ship$_{s1}$] that tried to [break$_{m2}$] Israels [blockade$_{s2}$] of [Gaza$_{loc1}$] [docked$_{e1}$] in the [Egyptian port of El Arish$_{loc2}$] on Thursday afternoon as the ships sponsor, a son of the Libyan leader, Col. Muammar el-Qaddafi, said that the [boat$_{s3}$] had [shifted course$_{m2}$] because the Israeli government agreed to allow Libya to support [building$_{e2}$] and [reconstruction$_{e3}$] in [Gaza$_{loc3}$].

The Libyan [ship$_{s4}$], the Amalthea, [docked$_{e4}$] five days after [setting off$_{m3}$] for [Gaza$_{loc4}$] from [Greece$_{loc5}$]. The [boat$_{s5}$] was [shadowed] and warned off by the [Israeli Navy$_{s6}$], and finally [headed] to [Egypt], where [it$_{s7}$] [lingered] for a day [just outside the port$_{loc6}$].

Figure 4: Spatially annotated text from newswire articles

derspecified except that it represents a particular change with respect to a previous motion; that is, the new direction is different from the old one. This suggests that to be complete, ISO-Space should allow the interpretation of motions to be dependent on other motions. This is currently not fleshed out in the specifications.

6 Corpus Driven Development

As ISO-Space aims to account for a wide range of spatial language phenomena, the development of a diverse set of corpora is crucial. At a recent gathering of the ISO-Space working group, five specific genres were examined as the specification was vetted and modified. The chosen genres were written directions, standard newswire, location descriptions, interior descriptions, and travel blogs. Members of the working group examined multiple selections from each genre with an eye towards improving the current specification. Excerpts from each genre are given below.

- **Written directions**: *Take I-66 West to Exit 43A (Gainesville/Warrenton) and proceed South on Rt. 29 for approximately 10 miles.*
- **Standard Newswire**: *A Libyan ship that tried to break Israels blockade of Gaza docked in the Egyptian port of El Arish on Thursday afternoon.*
- **Location descriptions**: *Times Square is a major commercial intersection in the borough of Manhattan in New York City, at the junction of Broadway and Seventh Avenue and stretching from West 42nd to West 47th Streets.*
- **Interior descriptions**: *Sitting on the top of the bookcase farthest from you is a potted plant that hangs down almost to the floor.*
- **Travel blogs**: *After spending a night with a family in Managua, my father and I biked for two days and then took a boat out to a volcanic island in the middle of Lake Nicaragua, Central Americas largest lake.*

Standard newswire has previously been the focus of a number of annotation efforts, including TimeML and SpatialML. Two examples of newswire text were examined (see Figure 4 for a few excerpts). The complexity of these news stories has had direct consequences on the ISO-Space specification (e.g., the need for a SPATIAL_ENTITY tag for entities such as *hurricanes*).

Another side effect of developing a spatially rich corpus for ISO-Space is the confirmation that there is a clear distinction between an annotation specification and an annotation guideline for a specific task. This contrast was most apparent with the inclusion of written directions. The working group looked at two examples of this kind of text.

The first, included in the examples above, came from a conference center website (the site of the working group meeting). Directions to the center from two different starting points were provided. The second example, which was created specifically for the meeting, gave directions from one location on a college campus to another. The tone of this example was informal, similar to what one might encounter in an e-mail.

While these examples had a wealth of explicit locations (e.g., city names, bridges, etc.), they also frequently contained language such as *leave the office and turn right twice* and *continue straight ahead for one mile*. However, dealing directly with imperative events in directions was deemed an issue for the annotation guidelines rather than the specification. Instead of adding complexity to the specification to account for this single genre, the goal of ISO-Space is to create a robust specification that can be used for a wide range of spatial language. Specific guidelines for dealing with written directions will make use of this specification by informing annotators how to use the elements of ISO-Space to account for imperative events. For example, they could be instructed to paraphrase the directions so that they are given in more of a narrative format.

By examining several different genres during the development of ISO-Space, the working group hopes to ensure that the specification is robust enough that it can be used for the annotation of many spatially rich corpora. The annotation guidelines will detail exactly how the specification should be used for each genre.

7 Conclusion and Outstanding Issues

In this paper, we have reported on work aimed at providing a comprehensive foundation for the annotation of spatial information in natural language text. While there are clearly many issues remaining, we have attempted to follow a strict methodology of specification development, as adopted by ISO TC37/SC4 and outlined in (Ide and Romary, 2004) and (Bunt and Romary, 2002), and as implemented with the development of ISO-TimeML and others in the family of SemAF standards.

Some of the issues which remain unaddressed in the current document include the following:

- Should orientation be a distinct relation type?
- Should there be a distinction between PATH and PATH-SEGMENTS, when referring to parts of a route, for example?
- What granularity or specificity of the qualitative spatial relations is appropriate?
- How should Goal and Intentional Locations be represented?
  a. John left New York for Boston.
  b. The plane is on its way to Paris.
- How do we provide for an expressive vocabulary for shapes of objects?
- Do we need a distinction between LOCATION and PLACE?
- Do we need SPATIAL-FUNCTIONS in order to refer to ad hoc regions, such as *behind the building*, *in front of the house*?

Many of these issues will most likely be resolved through the development of annotated corpus fragments, wherein alternative specification proposals can be tested for expressiveness, coverage, and ease of annotation. We are currently in the process of preparing such a corpus, while also developing the first fragment of a semantics for the abstract syntax presented here.

It should be noted that the selected corpora are all from written discourse. Clearly it would be beneficial to also consider other application areas including human-human or human-machine dialogues, for example car navigation systems, route planning, map searching, and navigating robots. In addition, all current corpora used are in English. In a later stage we will also include corpora from other languages since the ultimate idea is to have a multilingual annotation specification.

Acknowledgments

We would like to thank the members of the ISO-Space Working Group for their significant input to the current specification. Part of this research was funded under the NURI grant HM1582-08-1-0018 by the National Geospatial Agency.

References

John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence*.

Hans C. Boas. 2001. Frame semantics as a framework for describing polysemy and syntactic structures of english and german motion verbs in contrastive computational lexicography. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and

Shereen Khoja, editors, *Proceedings of the corpus linguistics 2001 conference*, volume 13, Lancaster, UK. University Center for Computer corpus research on language.

Harry Bunt and J. Pustejovsky. 2010. Annotating temporal and event quantification. In *Proceedings of 5th ISA Workshop*.

Harry Bunt and L. Romary. 2002. Proceedings of lrec 2002 workshop: Towards multimodal content representation.

Bert Cappelle and Renaat Declerck. 2005. Spatial and temporal boundedness in english motion events. *Journal of Pragmatics*, 37(6):889–917, June.

M.J. Egenhofer, E. Clementini, and P. Di Felice. 1994. Topological relations between regions with holes. *International Journal of Geographical Information Systems*, 8(2):129–142.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.

Andrew Frank. 1997. Spatial ontology: A geographical information point of view. In Oliviero Stock, editor, *Spatial and Temporal Reasoning*, pages 135–153. Springer Netherlands.

Pierre Grenon and Barry Smith. 2004. Snap and span: Towards dynamic spatial ontology. *Spatial Cognition and Computation*, 4(1):69–104.

John Herring, David M. Mark, John Herring (eds.), David M. Mark, Max J. Egenhofer, and Max J. Egenhofer. 1994. The 9-intersection: Formalism and its use for natural-language spatial predicates.

Annette Herskovits. 1986. *Language and Spatial Cognition: an Interdisci- plinary Study of the Prepositions in English*. Cambridge University Press.

Nancy Ide and L. Romary. 2003. Outline of the international standard linguistic annotation framework. In *Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right*.

N. Ide and L. Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3-4):211–225.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. 2006. Extensive classifications of english verbs. In *Proceedings of the 12th EURALEX International Congress*.

Yohei Kurata and Max Egenhofer. 2007. The 9+ intersection for topological relations between a directed line segment and a region. In B. Gottfried, editor, *Workshop on Behaviour and Monitoring Interpretation*, pages 62–76, Germany, September.

Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. 2008. Spatialml: Annotation scheme, corpora, and tools. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

George Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

MITRE. 2007. Spatialml: Annotation scheme for marking spatial expressions in natural language. http://sourceforge.net/projects/spatialml/, December.

Philippe Muller. 1998. A qualitative theory of motion based on spatio-temporal primitives. In Anthony G. Cohn, Lenhart Schubert, and Stuart C. Shapiro, editors, *KR'98: Principles of Knowledge Representation and Reasoning*, pages 131–141. Morgan Kaufmann, San Francisco, California.

James Pustejovsky and Jessica L. Moszkowicz. 2008. Integrating motion predicate classes with spatial and temporal annotations. In *Proceedings of COLING 2008*, Manchester, UK.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39:123–164, May.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*.

James Pustejovsky. 2006. Unifying linguistic annotations: A timeml case study. In *Proceedings of TSD 2006*, Brno, Czech Republic.

David Randell, Zhan Cui, and Anthony Cohn. 1992. A spatial logic based on regions and connections. In Morgan Kaufmann, editor, *Proceedings of the 3rd Internation Conference on Knowledge Representation and REasoning*, pages 165–176, San Mateo.

Leonard Talmy. 1983. How language structures space. In Herbert Pick and Linda Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*. Plenum Press.

An Annotation Scheme for Reichenbach's Verbal Tense Structure

Leon Derczynski and **Robert Gaizauskas**
Department of Computer Science
University of Sheffield, UK
{leon,robertg}@dcs.shef.ac.uk

Abstract

In this paper we present RTMML, a markup language for the tenses of verbs and temporal relations between verbs. There is a richness to tense in language that is not fully captured by existing temporal annotation schemata. Following Reichenbach we present an analysis of tense in terms of abstract time points, with the aim of supporting automated processing of tense and temporal relations in language. This allows for precise reasoning about tense in documents, and the deduction of temporal relations between the times and verbal events in a discourse. We define the syntax of RTMML, and demonstrate the markup in a range of situations.

1 Introduction

In his 1947 account, Reichenbach offered an analysis of the tenses of verbs, in terms of abstract time points. Reichenbach details nine tenses (see Table 1). The tenses detailed by Reichenbach are past, present or future, and may take a simple, anterior or posterior form. In English, these apply to single verbs and to verbal groups (e.g. *will have run*, where the main verb is *run*).

To describe a tense, Reichenbach introduces three abstract time points. Firstly, there is the speech time, S. This represents the point at which the verb is uttered or written. Secondly, event time E is the time that the event introduced by the verb occurs. Thirdly, there is reference time R; this is an abstract point, from which events are viewed. In Example 1, speech time S is when the author created the discourse (or perhaps when the reader interpreted it). Reference time R is *then* – an abstract point, before speech time, but after the event time E, which is the leaving of the building. In this sentence, one views events from a point in time later than they occurred.

(1) *By then, she had left the building.*

While we have rich annotation languages for time in discourse, such as TimeML[1] and TCNL[2], none can mark the time points in this model, or the relations between them. Though some may provide a means for identifying speech and event times in specific situations, there is nothing similar for reference times. All three points from Reichenbach's model are sometimes necessary to calculate the information used in these rich annotation languages; for example, they can help determine the nature of a temporal relation, or a calendrical reference for a time. We will illustrate this with two brief examples.

(2) *By April 26^{th}, it was all over.*

In Example 2, there is an anaphoric temporal expression describing a date. The expression is ambiguous because we cannot position it absolutely without an agreed calendar and a particular year. This type of temporal expression is interpreted not with respect to speech time, but with respect to reference time (Ahn et al., 2005). Without a time frame for the sentence (presumably provided earlier in the discourse), we cannot determine which year the date is in. If we are able to set bounds for R in this case, the time in Example 2 will be the April 26^{th} adjacent to or contained in R; as the word *by* is used, we know that the time is the April 26^{th} following R, and can normalise the temporal expression, associating it with a time on an absolute scale.

Temporal link labelling is the classification of relations between events or times. We might say an event of *the airport closed* occurred **after** another event of *the aeroplane landed*; in this case, we have specified the type of temporal relation between two events. This task is difficult to automate (Verhagen et al., 2010). There are clues in

[1] http://www.timeml.org; Boguraev et al. (2005).
[2] See Han et al. (2006).

discourse that human readers use to temporally relate events or times. One of these clues is tense. For example:

(3) *John told me the news, but I had already sent the letter.*

Example 3 shows a sentence with two verb events – *told* and *had sent*. Using Reichenbach's model, these share their speech time S (the time of the sentence's creation) and reference time R, but have different event times. In the first verb, reference and event time have the same position. In the second, viewed from when John told the news, the letter sending had already happened – that is, event time is before reference time. As reference time R is the same throughout the sentence, we know that the letter was sent before John mentioned the news. Describing S, E and R for verbs in a discourse and linking these points with each other (and with times) is the only way to ensure correct normalisation of all anaphoric and deictic temporal expressions, as well as enabling high-accuracy labelling of some temporal links.

Some existing temporal expression normalisation systems heuristically approximate reference time. GUTime (Mani and Wilson, 2000) interprets the reference point as "the time currently being talked about", defaulting to document creation date. Over 10% of errors in this system were directly attributed to having an incorrect reference time, and correctly tracking reference time is the only way to resolve them. TEA (Han et al., 2006) approximates reference time with the most recent time temporally before the expression being evaluated, excluding noun-modifying temporal expressions; this heuristic yields improved performance in TEA when enabled, showing that modelling reference time helps normalisation. HeidelTime (Strötgen and Gertz, 2010) uses a similar approach to TEA but does not exclude noun-modifying expressions.

The recently created WikiWars corpus of TIMEX2 annotated text prompted the comment that there is a "need to develop sophisticated methods for temporal focus tracking if we are to extend current time-stamping technologies" (Mazur and Dale, 2010). Resources that explicitly annotate reference time will be direct contributions to the completion of this task.

Elson and McKeown (2010) describe how to relate events based on a "perspective" which is calculated from the reference and event times of an event pair. They construct a natural language generation system that requires accurate reference times in order to correctly write stories. Portet et al. (2009) also found reference point management critical to medical summary generation.

These observations suggest that the ability to automatically determine reference time for verbal expressions is useful for a number of computational language processing tasks. Our work in this area – in which we propose an annotation scheme including reference time – is a first step in this direction.

In Section 2 we describe some crucial points of Reichenbach's model and the requirements of an annotation schema for tense in natural language. We also show how to reason about speech, event and reference times. Then, in Section 3, we present an overview of our markup. In Section 4 we give examples of annotated text (fictional prose and newswire text that we already have another temporal annotation for), event ordering and temporal expression normalisation. Finally we conclude in Section 5 and discuss future work.

2 Exploring Reichenbach's model

Each tensed verb can be described with three points; speech time, event time and reference time. We refer to these as S, E and R respectively. Speech time is when the verb is uttered. Event time is when the action described by the verb occurs. Reference time is a viewpoint from where the event is perceived. A summary of the relative positions of these points is given in Table 1.

While each tensed verb involves a speech, event and reference time, multiple verbs may share one or more of these points. For example, all narrative in a news article usually has the same speech time (that of document creation). Further, two events linked by a temporal conjunction (e.g. *after*) are very likely to share the same reference time.

From Table 1, we can see that conventionally English only distinguishes six tenses. Therefore, some English tenses will suggest more than one arrangement of S, E and R. Reichenbach's tense names suffer from this ambiguity too, but to a much lesser degree. When following Reichenbach's tense names, it is the case that for past tenses, R always occurs before S; in the future, R is always after S; and in the present, S and R are simultaneous. Further, "anterior" suggests E

Relation	*Reichenbach's Tense Name*	*English Tense Name*	*Example*
E<R<S	Anterior past	Past perfect	*I had slept*
E=R<S	Simple past	Simple past	*I slept*
R<E<S	Posterior past		*I expected that ..*
R<S=E			*I would sleep*
R<S<E			
E<S=R	Anterior present	Present perfect	*I have slept*
S=R=E	Simple present	Simple present	*I sleep*
S=R<E	Posterior present	Simple future	*I will sleep (Je vais dormir)*
S<E<R	Anterior future	Future perfect	*I will have slept*
S=E<R			
E<S<R			
S<R=E	Simple future	Simple future	*I will sleep (Je dormirai)*
S<R<E	Posterior future		*I shall be going to sleep*

Table 1: Reichenbach's tenses; from Mani et al. (2005)

before R, "simple" that R and E are simultaneous, and "posterior" that E is after R. The flexibility of this model permits the full set of available tenses (Song and Cohen, 1988), and this is sufficient to account for the observed tenses in many languages.

Our goal is to define an annotation that can describe S, E and R (speech, event and reference time) throughout a discourse. The lexical entities that these times are attached to are verbal events and temporal expressions. Therefore, our annotation needs to locate these entities in discourse, and make the associated time points available.

2.1 Special properties of the reference point

The reference point R has two special uses. When sentences or clauses are combined, grammatical rules require tenses to be adjusted. These rules operate in such a way that the reference point is the same in all cases in the sequence. Reichenbach names this principle **permanence of the reference point**.

Secondly, when temporal expressions (such as a TimeML TIMEX3 of type DATE, but not DURATION) occur in the same clause as a verbal event, the temporal expression does not (as one might expect) specify event time E, but instead is used to position reference time R. This principle is named **positional use of the reference point**.

2.2 Context and the time points

In the linear order that events and times occur in discourse, speech and reference points persist until changed by a new event or time. That is, the reference time from one sentence will roll over to the next sentence, until it is repositioned explicitly by a tensed verb or time. To cater for subordinate clauses in cases such as reported speech, we add a caveat – S and R persist as a discourse is read in textual order, for each context. We can define a context as an environment in which events occur, such as the main body of the document, reported speech, or the conditional world of an *if* clause (Hornstein, 1990). For example:

(4) *Emmanuel had said "This will explode!", but changed his mind.*

Here, *said* and *changed* share speech and reference points. Emmanuel's statement occurs in a separate context, which the opening quote instantiates, ended by the closing quote (unless we continue his reported speech later), and begins with an S that occurs at the same time as *said*'s E. This persistence must be explicitly stated in RTMML.

2.3 Capturing the time points with TimeML

TimeML is a rich, developed standard for temporal annotation. There exist valuable resources annotated with TimeML that have withstood significant scrutiny. However TimeML does not address the issue of annotating Reichenbach's tense model with the goal of understanding reference time or creating resources that enable detailed examination of the links between verbal events in discourse.

Although TimeML permits the annotation of tense for `<EVENT>`s, it is not possible to unambiguously map its tenses to Reichenbach's model. This restricts how well we can reason about verbal

events using TimeML-annotated documents. Of the usable information for mapping TimeML annotations to Reichenbach's time points, TimeML's `tense` attribute describes the relation between S and E, and its `aspect` attribute can distinguish between `PERFECTIVE` and `NONE` – that is, between $E < R$ and a conflated class of $(E = R) \vee (R < E)$. Cases where $R < E$ are often awkward in English (as in Table 1), and may even lack a distinct syntax; the French *Je vais dormir* and *Je dormirai* both have the same TimeML representation and both translate to *I will sleep* in English, despite having different time point arrangements.

It is not possible to describe or build relations to reference points at all in TimeML. It may be possible to derive the information about S, E and R directly represented in our scheme from a TimeML annotation, though there are cases – especially outside of English – where it is not possible to capture the full nuance of Reichenbach's model using TimeML. An RTMML annotation permits simple reasoning about reference time, and assist the labelling of temporal links between verb events in cases where TimeML's tense and aspect annotation is insufficient. This is why we propose an annotation, and not a technique for deriving S, E, and R from TimeML.

3 Overview of RTMML

The annotation schema RTMML is intended to describe the verbal event structure detailed in Reichenbach (1947), in order to permit the relative temporal positioning of reference, event, and speech times. A simple approach is to define a markup that only describes the information that we are interested in, and can be integrated with TimeML. For expositional clarity we use our own tags but it is possible (with minor modifications) to integrate them with TimeML as an extension to the standard.

Our procedure is as follows. Mark all times and verbal events (e.g. TimeML TIMEX3s and those EVENTs whose lexical realisation is a verb) in a discourse, as $T_1..T_n$ and $V_1..V_n$ respectively. We mark times in order to resolve positional uses of the reference point. For each verbal event V_i, we may describe or assign three time points S_i, E_i, and R_i. Further, we will relate T, S, E and R points using disjunctions of the operators $<$, $=$ and $>$. It is not necessary to define a unique set of these points for each verb – in fact, linking them across a discourse helps us temporally order events and track reference time. We can also define a "discourse creation time," and call this S_D.

(5) *John said, "Yes, we have left".*

If we let *said* be V_1 and *left* be V_2:

- $S_1 = S_D$

From the tense of V_1 (simple past), we can say:

- $R_1 = S_1$
- $E_1 < R_1$

As V_2 is reported speech, it is true that:

- $S_2 = E_1$

Further, as V_2 is anterior present:

- $R_2 = S_2$
- $E_2 < R_2$

As the $=$ and $<$ relations are transitive, we can deduce an event ordering $E_2 < E_1$.

3.1 Annotation schema

The annotation language we propose is called RTMML, for Reichenbach Tense Model Markup Language. We use standoff annotation. This keeps the text uncluttered, in the spirit of *ISO LAF* and *ISO SemAF-Time*. Annotations reference tokens by index in the text, as can be seen in the examples below. Token indices begin from zero. We explicitly state the segmentation plan with the `<seg>` element, as described in Lee and Romary (2010) and *ISO DIS 24614-1 WordSeg-1*.

The general speech time of a document is defined with the `<doc>` element, which has one or two attributes: an ID, and (optionally) `@time`. The latter may have a normalised value, formatted according to TIMEX3 (Boguraev et al., 2005) or TIDES (Ferro et al., 2005), or simply be the string `now`.

Each `<verb>` element describes a tensed verbal group in a discourse. The `@target` attribute references token offsets; it has the form `target="#token0"` or `target="#range(#token7,#token10)"` for a 4-token sequence. Comma-separated lists of offsets are valid, for situations where verb groups are non-contiguous. Every `verb` has a unique value in its `@id` attribute. The tense of a verb group is described using the attributes `@view`

Relation name	Interpretation
POSITIONS	$T_a = R_b$
SAME_TIMEFRAME	$R_a = R_b[, R_c, ..R_x]$
REPORTS	$E_a = S_b$

Table 2: The meaning of a certain link type between verbs or times a and b.

(with values *simple*, *anterior* or *posterior*) and `@tense` (*past*, *present* or *future*).

The `<verb>` element has optional `@s`, `@e` and `@r` attributes; these are used for directly linking a verb's speech, event or reference time to a time point specified elsewhere in the annotation. One can reference document creation time with a value of `doc` or a temporal expression with its id (for example, `t1`). To reference the speech, event or reference time of other verbs, we use hash references to the event followed by a dot and then the character `s`, `e` or `r`; e.g., `v1`'s reference time is referred to as `#v1.r`.

As every tensed verb always has exactly one S, E and R, and these points do not hold specific values or have a position on an absolute scale, we do not attempt to directly annotate them or place them on an absolute scale. One might think that the relations should be expressed in XML links; however this requires reifying time points when the information is stored in the relations between time points, so we focus on the relations between these points for each `<verb>`. To capture these internal relations (as opposed to relations between the S, E and R of different verbs), we use the attributes `se`, `er` and `sr`. These attributes take a value that is a disjunction of $<$, $=$ and $>$.

Time-referring expressions are annotated using the `<timerefx>` element. This has an `@id` attribute with a unique value, and a `@target`, as well as an optional `@value` which works in the same was as the `<doc>` element's `@time` attribute.

```
<rtmml>
Yesterday, John ate well.
 <seg type="token" />
 <doc time="now" />
 <timerefx xml:id="t1" target="
       #token0" />
 <verb xml:id="v1" target="#token3"
       view="simple" tense="past"
       sr=">" er="=" se=">"
       r="t1" s="doc" />
 </rtmml>
```

In this example, we have defined a time *Yesterday* as `t1` and a verbal event *ate* as `v1`. We have categorised the tense of `v1` within Reichenbach's nomenclature, using the `verb` element's `@view` and `@tense` attributes.

Next, we directly describe the reference point of `v1`, as being the same as the time `t1`. Finally, we say that this verb is uttered at the same time as the whole discourse – that is, $S_{v1} = S_D$. In RTMML, if the speech time of a verb is not otherwise defined (directly or indirectly) then it is S_D. In cases of multiple voices with distinct speech times, if a speech time is not defined elsewhere, a new one may be instantiated with a string label; we recommend the formatting *s*, *e* or *r* followed by the verb's ID.

This sentence includes a positional use of the reference point, annotated in `v1` when we say `r="t1"`. To simplify the annotation task, and to verbosely capture a use of the reference point, RTMML permits an alternative annotation with the `<rtmlink>` element. This element takes as arguments a relation and a set of times and/or verbs. Possible relation types are POSITIONS, SAME_TIMEFRAME (annotating permanence of the reference point) and REPORTS for reported speech; the meanings of these are given in Table 2. In the above markup, we could replace the `<verb>` element with the following:

```
<verb xml:id="v1" target="#token3"
      view="simple" tense="past"
      sr=">" er="=" se=">" s="doc" />
<rtmlink xml:id="l1" type="POSITIONS">
  <link source="#t1" />
  <link target="#v1" />
</rtmlink>
```

When more than two entities are listed as targets, the relation is taken as being between an optional `source` entity and each of the `target` entities. Moving inter-verbal links to the `<rtmlink>` element helps fulfil *TEI p5* and the *LAF* requirements that referencing and content structures are separated. Use of the `<rtmlink>` element is not compulsory, as not all instances of positional use or permanence of the reference point can be annotated using it; Reichenbach's original account gives an example in German.

3.2 Reasoning and inference rules

Our three relations $<$, $=$ and $>$ are all transitive. A minimal annotation is acceptable. The S, E and R points of all verbs, S_D and all Ts can represent nodes on a graph, connected by edges labelled

with the relation between nodes.

To position all times in a document with maximal accuracy, that is, to label as many edges in such a graph as possible, one can generate a closure by means of deducing relations. An agenda-based algorithm is suitable for this, such as the one given in Setzer et al. (2005).

3.3 Integration with TimeML

To use RTMML as an ISO-TimeML extension, we recommend that instead of annotating and referring to <timerefx>s, one refers to <TIMEX3> elements using their tid attribute; references to <doc> will instead refer to a <TIMEX3> that describes document creation time. The attributes of <verb> elements (except xml:id and target) may be be added to <MAKEINSTANCE> or <EVENT> elements, and <rtmlink>s will refer to event or event instance IDs.

4 Examples

In this section we will give developed examples of the RTMML notation, and show how it can be used to order events and position events on an external temporal scale.

4.1 Annotation example

Here we demonstrate RTMML annotation of two short pieces of text.

4.1.1 Fiction

From *David Copperfield* by Charles Dickens:

(6) *When he had put up his things for the night he took out his flute, and blew at it, until I almost thought he would gradually blow his whole being into the large hole at the top, and ooze away at the keys.*

We give RTMML for the first five verbal events from Example 6 RTMML in Figure 1. The fifth, v5, exists in a context that is instantiated by v4; its reference time is defined as such. We can use one link element to show that v2, v3 and v4 all use the same reference time as v1. The temporal relation between event times of v1 and v2 can be inferred from their shared reference time and their tenses; that is, given that v1 is anterior past and v2 simple past, we know $E_{v1} < R_{v1}$ and $E_{v2} = R_{v2}$. As our <rtmlink> states $R_{v1} = R_{v2}$, then $E_{v1} < E_{v2}$. Finally, v5 and v6 happen in the same context, described with a second SAME_TIMEFRAME link.

4.1.2 Editorial news

From an editorial piece in TimeBank (Pustejovsky et al., 2003) (AP900815-0044.tml):

(7) *Saddam appeared to accept a border demarcation treaty he had rejected in peace talks following the August 1988 cease-fire of the eight-year war with Iran.*

```
<doc time="1990-08-15T00:44" />
<!-- appeared -->
<verb xml:id="v1" target="#token1"
  view="simple" tense="past" />
<!-- had rejected -->
<verb xml:id="v2"
  target="#range(#token9,#token10)"
  view="anterior" tense="past" />
<rtmlink xml:id="l1"
  type="SAME_TIMEFRAME">
  <link target="#v1" />
  <link target="#v2" />
</rtmlink>
```

Here, we relate the simple past verb *appeared* with the anterior past (past perfect) verb *had rejected*, permitting the inference that the first verb occurs temporally after the second. The corresponding TimeML (edited for conciseness) is:

```
Saddam <EVENT eid="e74" class="I_STATE">
appeared</EVENT> to accept a border
demarcation treaty he had <EVENT eid="e77"
class="OCCURRENCE">rejected</EVENT>

<MAKEINSTANCE eventID="e74" eiid="ei1568"
 tense="PAST" aspect="NONE" polarity="POS"
 pos="VERB"/>
<MAKEINSTANCE eventID="e77" eiid="ei1571"
 tense="PAST" aspect="PERFECTIVE"
 polarity="POS" pos="VERB"/>
```

In this example, we can see that the TimeML annotation includes the same information, but a significant amount of other annotation detail is present, cluttering the information we are trying to see. Further, these two <EVENT> elements are not directly linked, requiring transitive closure of the network described in a later set of <TLINK> elements, which are omitted here for brevity.

4.2 Linking events to calendrical references

RTMML makes it possible to precisely describe the nature of links between verbal events and times, via positional use of the reference point. We will link an event to a temporal expression, and suggest a calendrical reference for that expression, allowing the events to be placed on a calendar. Consider the below text, from wsj_0533.tml in TimeBank.

```
<doc time="1850" mod="BEFORE" />
<!-- had put -->
<verb xml:id="v1"
  target="#range(#token2,#token3)"
  view="anterior" tense="past" />
<!-- took -->
<verb xml:id="v2" target="#token11"
  view="simple" tense="past" />
<!-- blew -->
<verb xml:id="v3" target="#token17"
  view="simple" tense="past" />
<!-- thought -->
<verb xml:id="v4" target="#token24"
  view="simple" tense="past" />
<!-- would gradually blow -->
<verb xml:id="v5"
  target="#range(#token26,#token28)"
  view="posterior" tense="past"
  se="=" er=">" sr=">"
  r="#v4.e" />
<!-- ooze -->
<verb xml:id="v6"
  target="#range(#token26,#token28)"
  view="posterior" tense="past"
  se="=" er=">" sr=">" />
<rtmlink xml:id="l1"
  type="SAME_TIMEFRAME">
  <link target="#v1" />
  <link target="#v2" />
  <link target="#v3" />
  <link target="#v4" />
</rtmlink>
<rtmlink xml:id="l2"
  type="SAME_TIMEFRAME">
  <link target="#v5" />
  <link target="#v6" />
</rtmlink>
```

Figure 1: RTMML for a passage from David Copperfield.

(8) *At the close of business Thursday, 5,745,188 shares of Connaught and C$44.3 million face amount of debentures, convertible into 1,826,596 common shares, had been tendered to its offer.*

```
<doc time="1989-10-30" />
<!-- close of business Thursday -->
<timerefx xml:id="t1"
  target="#range(#token2,#token5)" />
<!-- had been tendered -->
<verb xml:id="v1"
  target="#range(#token25,#token27)"
  view="anterior" tense="past" />
<rtmlink xml:id="l1" target="#t1 #v1">
  <link target="#t1" />
  <link target="#v1" />
</rtmlink>
```

This shows that the reference time of `v1` is `t1`. As `v1` is anterior, we know that the event mentioned occurred before *close of business Thursday*. Normalisation is not a task that RTMML addresses, but there are existing methods for deciding which Thursday is being referenced given the document creation date (Mazur and Dale, 2008); a time of day for *close of business* may be found in a gazetteer.

4.3 Comments on annotation

As can be seen in Table 1, there is not a one-to-one mapping from English tenses to the nine specified by Reichenbach. In some annotation cases, it is possible to see how to resolve such ambiguities. Even if view and tense are not clearly determinable, it is possible to define relations between S, E and R; for example, for arrangements corresponding to the simple future, $S < E$. In cases where ambiguities cannot be resolved, one may annotate a disjunction of relation types; in this example, we might say "$S < R$ or $S = R$" with `sr="<="`.

Contexts seem to have a shared speech time, and the $S - R$ relationship seems to be the same throughout a context. Sentences which contravene this (e.g. *"By the time I ran, John will have arrived"*) are rather awkward.

RTMML annotation is not bound to a particular language. As long as a segmentation scheme (e.g. WordSeg-1) is agreed and there is a compatible system of tense and aspect, the model can be applied and an annotation created.

5 Conclusion and Future Development

Being able to recognise and represent reference time in discourse can help in disambiguating temporal reference, determining temporal relations between events and in generating appropriately tensed utterances. A first step in creating computational tools to do this is to develop an annotation schema for recording the relevant temporal information in discourse. To this end we have presented RTMML, our annotation for Reichenbach's model of tense in natural language.

We do not intend to compete with existing languages that are well-equipped to annotate temporal information in documents; RTMML may be integrated with TimeML. What is novel in RTMML is the ability to capture the abstract parts of tense in language. We can now annotate Reichenbach's time points in a document and then process them, for example, to observe interactions between temporal expressions and events, or to track reference time through discourse. This is not directly possible with existing annotation languages.

There are some extensions to Reichenbach's model of the tenses of verbs, which RTMML does not yet cater for. These include the introduction of a reference interval, as opposed to a reference point, from Dowty (1979), and Comrie's suggestion of a second reference point in some circumstances (Comrie, 1985). RTMML should cater for these extensions.

Further, we have preliminary annotation tools and have begun to create a corpus of annotated texts that are also in TimeML corpora. This will allow a direct evaluation of how well TimeML can represent Reichenbach's time points and their relations. To make use of Reichenbach's model in automatic annotation, given a corpus, we would like to apply machine learning techniques to the RTMML annotation task. Work in this direction should enable us to label temporal links and to anchor time expressions with complete accuracy where other systems have not succeeded.

6 Acknowledgements

The authors would like to thank David Elson for his valuable comments. The first author would also like to acknowledge the UK Engineering and Physical Science Research Council's support in the form of a doctoral studentship.

References

D. Ahn, S.F. Adafre, and MD Rijke. 2005. Towards task-based temporal extraction and recognition. In *Dagstuhl Seminar Proceedings*, volume 5151.

B. Boguraev, J. Castano, R. Gaizauskas, B. Ingria, G. Katz, B. Knippen, J. Littman, I. Mani, J. Pustejovsky, A. Sanfilippo, et al. 2005. TimeML 1.2. 1: A Formal Specification Language for Events and Temporal Expressions.

B. Comrie. 1985. *Tense*. Cambridge University Press.

D.R. Dowty. 1979. *Word meaning and Montague grammar*. Kluwer.

D. Elson and K. McKeown. 2010. Tense and Aspect Assignment in Narrative Discourse. In *Proceedings of the Sixth International Conference in Natural Language Generation*.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. TIDES 2005 standard for the annotation of temporal expressions. Technical report, MITRE.

International Organization for Standardization. 2009a. *ISO DIS 24612 LRM - Language Annotation Framework (LAF)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization. 2009b. *ISO DIS 24614-1 LRM - Word Segmentation of Text - Part 1: Basic Concepts and General Principles (WordSeg-1)*. ISO/TC 37/SC 4/WG 2.

International Organization for Standardization. 2009c. *ISO DIS 24617-1 LRM - Semantic Annotation Framework - Part 1: Time and Events (SemAF-Time)*. ISO/TC 37/SC 4/WG 2.

B. Han, D. Gates, and L. Levin. 2006. From language to time: A temporal expression anchorer. In *Temporal Representation and Reasoning (TIME)*, pages 196–203.

N. Hornstein. 1990. *As time goes by: Tense and universal grammar*. MIT Press.

K. Lee and L. Romary. 2010. Towards Interoperability of ISO Standards for Language Resource Management. In *International Conference on Global Interoperability for Language Resources*.

I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on ACL*, pages 69–76. ACL.

I. Mani, J. Pustejovsky, and R. Gaizauskas. 2005. *The language of time: a reader*. Oxford University Press, USA.

P. Mazur and R. Dale. 2008. What's the date?: high accuracy interpretation of weekday names. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 553–560. ACL.

P. Mazur and R. Dale. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922. ACL.

F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

J. Pustejovsky, P. Hanks, et al. 2003. The TimeBank corpus. In *Corpus Linguistics*, volume 2003, page 40.

H. Reichenbach. 1947. The Tenses of Verbs. *Elements of Symbolic Logic*, pages 287–98.

A. Setzer, R. Gaizauskas, and M. Hepple. 2005. The role of inference in the temporal annotation and analysis of text. *Language Resources and Evaluation*, 39(2):243–265.

F. Song and R. Cohen. 1988. The interpretation of temporal relations in narrative. In *Proceedings of the 7th National Conference of AAAI*.

J. Strötgen and M. Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th Workshop on Semantic Evaluation*, pages 321–324. ACL.

M. Verhagen, R. Saurí, T. Caselli, and J. Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th Workshop on Semantic Evaluation*, pages 57–62. ACL.

Multi-level discourse relations between dialogue units

Volha Petukhova
Tilburg Center for Creative Computing
Tilburg University, The Netherlands,
v.petukhova@uvt.nl

Laurent Prévot
Laboratoire Parole et Langage
Université de Provence & CNRS
laurent.prevot@lpl-aix.fr

Harry Bunt
Tilburg Center for Creative Computing
Tilburg University, The Netherlands,
harry.bunt@uvt.nl

Abstract

In order to analyse what happens in dialogue it is insufficient to consider the content of its segments in isolation. In this paper we propose a meta-model for integrating discourse relations into a standard framework of dialogue act annotation (Bunt et al., 2010), considering in particular the various dialogue units involved and the nature of their relations.

1 Introduction

In discourse modelling, we need dialogue units and relations between them. This is uncontroversial, but the nature, the purpose and the definitions of units in discourse and their relations are the subject of much controversy (see e.g. Hovy, 1990). To the rhetorical relations identified in monologue (e.g explanation, justification, cause,...), dialogue adds relations such as those between a question and an answer, and between an utterance and feedback about its understanding.

Many frameworks for discourse analysis have attempted to capture discourse coherence by integrating all discourse segments into a single structure thanks to discourse relations. Although this has not always been made explicit, the assumption that there is a single ”coherence” dimension is strong in many frameworks (Hobbs, 1985; Mann and Thompson, 1988; Asher and Lascarides, 2003). Grosz and Sidner (1986), followed by Moore and Pollack (1992), on the other hand argued for the interplay between several structures to explain discourse phenomena. Petukhova and Bunt (2009) have shown that discourse markers are in general multifunctional, thus requiring a multidimensional approach.

In this paper we propose a meta-model for integrating various types of discourse relations into a standard framework of dialogue act annotation (Bunt et al., 2010), considering in particular the various dialogue units involved and the nature of their relations.

This paper is organized as follows. First, we briefly review the literature on discourse structure (Section 1.1). We describe the relevant aspects of the semantic framework that we will use to study relations between different types of dialogue units in Section 2. Section 3 discusses dialogue units, while the semantic relations between them are discussed in Section 4. Section 5 presents an empirical analysis of the scope of different types of semantic discourse relations and of the distances between related segments in two different types of dialogue. Section 6 concludes by summarizing the analysis in a meta-model and outlining perspectives for further research.

1.1 Previous work on discourse structure

A variety of frameworks for modelling discourse structure have been proposed since (Hobbs, 1979). While Van Dijk (1979) and Polanyi (1998) have attempted a quasi-syntactic approach, most frameworks are more functional and rely on interpretation for deriving a structure of discourse. Relations between discourse segments have in these frameworks been divided into several categories: semantic/ inter-propositional/ ideational/ content-level/ information-level; pragmatic/ intentional/ cognitive/ speech-act; presentational/ structural/ textual; see Hovy et al. (1995) for a discussion of the different categories.

Discourse relations can apply to segments of various size, from syntactic clauses to paragraphs. When considering dialogue, the picture gets even more complicated, with units specific to their interactive nature, such as speech-turns. Some researchers distinguish between macro-, meso- and micro-levels in discourse structuring (e.g. (Nakatani and Traum, 1992) and (Louwerse and Mitchell, 2003)), where the *micro-level* is con-

cerned with relations within a turn or within a single utterance; the *meso-level* concerns relations involving complete contributions in Clark's sense (Clark, 1996), typically an initiative and a reactive, corresponding to grounding units; and the *macro-level* is concerned with entire subdialogues, topic structure and elements of a plan-based analysis.

Although often cited as a crucial issue for linguistics and NLP, discourse structure frameworks face the problem of their empirical validation. It is mainly to address this issue that several discourse annotation projects have been undertaken in recent years (Carlson et al., 2001; Wolf and Gibson, 2005; Miltsakaki et al., 2004; Reese et al., 2007; Stede, 2008; Prasad et al., 2008). These ambitious projects share a common goal but differ greatly with regard to their theoretical assumptions. A more generic approach to the analysis of these relations would therefore be of great help for comparing and perhaps combining these accounts.

2 Semantic framework

Participants in dialogue produce utterances in order to provoke change in their addressees, and dialogue utterances can therefore be viewed as actions with intended state-changing effects on 'information states' (Poesio and Traum, 1998; Larsson and Traum, 2000; Bunt, 2000). Such communicative actions are called *dialogue acts*, and have two main components: a *semantic* (referential, propositional, or action-related) *content* and a *communicative function*, which specifies how an addressee is intended to update his information state with the semantic content.

In this study we use the semantic framework of Dynamic Interpretation Theory (DIT, Bunt, 2000), which takes a multidimensional view on dialogue in the sense that participation in a dialogue is viewed as performing several activities in parallel, such as pursuing the dialogue task, providing and eliciting feedback, and taking turns.

The DIT framework supports a 'multidimensional' semantics by relating context update operators to different compartments of structured context models which include, besides information states of the usual kind (beliefs and goals related to a task domain), also a dialogue history, information about the agent's processing state, beliefs about the dialogue partners' processing states, information and goals concerning the allocation of turns, and so on, relating to the various 'dimensions' that dialogue acts belong to. The interpretation of a multifunctional stretch of communicative behaviour corresponds to updating the context models of the communicating agents in multiple ways, combining the effects of each of the component functions.

3 Units in dialogue

The assignment of relations between certain units in dialogue presupposes a way to identify such units. Spoken dialogues are traditionally segmented into *turns*, stretch of communicative behaviour produced by one speaker, bounded by periods of inactivity of that speaker or by activity of another speaker (Allwood, 1992). The idea that a dialogue can be cut up into sequences of communicative activity of different speakers does not do justice to the complexity of natural communication, especially when more than two participants are involved. In natural communication, where the participants do not only use speech to communicate but also nonverbal and vocal signals (laughs, sighs, chuckles,..), all participants are most of the time performing *some* communicative activity. We therefore use a somewhat different notion, called a *turn unit*. A turn unit is a stretch of communicative behaviour produced by one participant which includes the use of speech, and is bounded by periods where that participant is not speaking. According to this definition a turn unit is produced by a speaker who may, in addition to speaking, also produce nonverbal communicative behaviour (such as gestures and facial expressions), and turn units produced by different speakers may overlap.

Turn units are relevant for the analysis of discourse relations in dialogue. Relations between turn units and other dialogue units occur when the speaker addresses an incomplete dialogue act, a discontinuous dialogue act, or parts of a dialogue act that is spread over several turn units. For example[1]:

(1) G1: you are going
F2: mmhm
G2: straight north

With F2, the speaker provides feedback on the turn unit G1, the first part of the dialogue act *you are going straight north*.

Turn units may consist of several *utterances*, linguistically defined stretches of communicative behaviour that have one or more communicative

[1] From the MapTask dialogue corpus.

functions. The stretches of behaviour that are relevant for interpretation as dialogue acts often coincide with utterances in this sense, but they may be discontinuous, may overlap, and may even contain parts of more than one turn. Communicative functions can be assigned more accurately to smaller units, called *functional segments*, which are defined as the functionally relevant minimal stretches of communicative behaviour (Geertzen et al., 2007). For example:

(2) A: Is that your opinion too, Bert?
B: Ehm,.. well,... I guess so.

In the first turn unit of (2), speaker A asks a question to B and assigns the turn to B (by the combined use of B's name, the intonation, and by looking at B). Speaker B performs a stalling act in order to buy some time for deciding what to say; now the fact that he starts speaking without waiting until he has made up his mind about his answer indicates that he accepts the turn. So the segment *Ehm,.. well,...* has both a stalling function and a turn-accepting function.

We also find relations in dialogue between functional segments and *groups of functional segments*, as the following example shows[2]:

(3) G1: Right. Start off facing north, turn to your left and walk forward, then to your left again. Keep walking forward until you come to the site of a plane crash. Go right roundabout it and turn to your right, so you end up facing north again.
U1: Could you just slow down a bit please?
G2: Sorry.
G3: So you start facing north
U2: Mmhmm

U1 is a negative feedback act relating to the group of 4 functional segments in G1. The speaker in U1 is apparently overloaded with the information given in G1, making it hard for him to process these segments successfully.

We also find relations between dialogue acts (more about that in the next section), as well as between a dialogue act and *group of dialogue acts*. When the dialogue acts in such a group are all concerned with a particular discussion topic, subdialogue or sub-task, such a group is often called a *discourse unit*.[3] An example is given in (18), where the utterance D2, requesting to recapitulate the discussion, by implication has a negative feedback function related to the discourse unit B1 – B5.

Example (4) shows a relation between a dialogue act and a group of dialogue acts which is smaller than a discourse unit.

(4) A1: can you tell me what time is the first flight in the morning to Munich?
B1: on what day do you want to travel?
A2: tomorrow.
B2: tomorrow morning
B3: the first flight that I have is at 7:45.

The dialogue act in B3 has functionally related with the group consisting of the question in A1 and the answer (to B1) in A2, which together are equivalent to a more complete question which B3 answers.

4 Relations between dialogue units

4.1 Functional and feedback dependence relations

Responsive dialogue acts by their very nature depend for their semantic content on the semantic content of the dialogue acts that they respond to. Responsive dialogue acts (also known as 'backwards-looking acts') come in three types:

A acts with a responsive general-purpose communicative function: Answer and its specializations (Confirm, Disconfirm, Correction); Agreement and Disagreement; and Address/Accept/Decline Request, Suggestion, or Offer;

B feedback acts with a dimension-specific communicative function;

C some dialogue acts with dimension-specific functions other than feedback functions, such as Return Greeting, Return Self-Introduction. Accept Apology, Accept Thanking, Return Goodbye; Completion and Correct-Misspeaking; and Turn Accept.

All responsive dialogue acts have a 'functional antecedent', being the dialogue acts that they respond to; those of type A have a semantic content that is co-determined by that of their functional antecedent. This relation between two dalogue acts (or between a dialogue act and a group of dialogue acts, as in (4)) is called a *functional dependence relation* (Bunt et al., 2010); it is a relation between the semantic contents of two dialogue acts that is due to their communicative functions. More explicitly:

[2] From the MapTask dialogue corpus.

[3] This is not the same as the notion of 'discourse unit' proposed by (Traum, 1994) for describing grounding in dialogue, which consists of an initial presentation and as many utterances as needed to make the initial utterance mutually understood. The two notions sometimes coincide, but not in general.

(5) *A functional dependence relation exists between a dialogue act DA_1 and one or more previous dialogue acts $\{DA_N\}$ if, due to the responsive communicative function of DA_1, the determination of its semantic content requires the semantic content of $\{DA_N\}$.*

An example of a functional dependence relation is (6), where the interpretation of A1 clearly depends very much on whether it is an answer to the question B1 or to the question B2 even though A1 would seem a complete, self-contained utterance.

(6) A1: I'm expecting Jan, Alex, Claudia, and David, and maybe Olga and Andrei to come.
B1: Do you know who's coming tonight?
B2: Which of the project members d'you think will be there?

Responsive dialogue acts of type B provide or elicit information about the (perceived) success in processing a segment of communicative behaviour earlier in the dialogue. We call this relation a *feedback dependence relation*. More explicitly:

(7) *A feedback dependence relation is a relation between a feedback act and the stretch of communicative behaviour whose processing the act provides or elicits information about.*

Examples are the relation between F2 and G11 in (1); between U1 and G1 in (3); and between B1 and A1 in (4).

Feedback acts refer explicitly or implicitly to the stretch of dialogue that they provide or elicit information about. This stretch of dialogue forms part of its semantic content. For example, the semantic content of the feedback act in F2 in (1), where the communicative function is Auto-Interpretation Positive,[4] has the segment G1 as its semantic content. In view of this relation between the feedback act and its functional antecedent, one could consider the feedback dependence relation as an instance of he functional dependence relation in the feedback dimension. However, the two relations must be distinguished, since a dialogue act with a functional dependence relation *also*, by implication, has a feedback dependence relation to its functional antecedent. For example, an answer implies positive feedback about the speaker's processing of the utterance expressing the question that the answer functionally depends on.

A feedback act does not necessarily refer to a single utterance, but may also relate to a larger stretch of dialogue; even to the entire preceding dialogue, like the global positive feedback expressed by *Okay* just before ending a dialogue. The scope and distance that may be covered by the various kinds of relations in discourse are analysed in the next section.

A responsive act of type C relates two dialogue acts, like in case A, but since these types of dialogue acts have no or only marginal semantic content, this does not lead to a semantically important relation, and we will disregard it in this paper.

4.2 Rhetorical relations

Rhetorical relations have been proposed as an explanation for the construction of coherence in discourse or at least as crucial modelling tools for capturing this coherence (Hobbs, 1985; Mann and Thompson, 1988; Asher and Lascarides, 2003; Sanders et al., 1992). The idea is that two text segments or sentences in written discourse, or two segments or utterances in dialogue, are linked together by means of certain relations, for which various terms have been used such as 'rhetorical relations', 'coherence relations', or 'discourse relations'.

Their study can be traced back to the Antiquity, with a continuous attention from rhetorics over the centuries, but the way they have been used recently in AI and NLP probably comes from Hobbs' seminal work in this area (Hobbs, 1979). Since then a range of taxonomies have been proposed in the literature to define relations in discourse. The well-known set of relations and their organization proposed by Mann and Thompson (1988), forming the core of Rhetorical Structure Theory, consists of 23 relations. This set is not claimed to be exhaustive, however. Hovy and Maier (1995) studied approximately 30 alternative proposals and fused the various taxonomies into more than 400 relations. They proposed a hierarchical taxonomy of approximately 70 relations.

Some rhetorical relations, such as *Explanation, Justification*, and *Cause* are clearly semantic, whereas others, like *First, Second,... , Finally; Concluding* are more presentational in nature. The occurrence of truly semantic rhetorical relations is illustrated in example (8) from the AMI corpus[5], where participant A talks about remote controls:

(8) A1: You keep losing them
A2: It slips behind the couch or it's kicked under the table

[4] This is the communicative function expressing that the speaker informs the addressee that he (believes that he) understands the utterance that the feedback is about.

[5] See `http://www.amiproject.org/`

The events described in these sentences are semantically related by *Cause* relations: *Cause* (slipped; keep_loosing) and *Cause* (kicked; keep_loosing). In cases like this the two sentences are related through a rhetorical relation between the events they contain. In this paper we will use the term 'inter-propositional relation' for rhetorical relations between the semantic contents of two dialogue acts, irrespective of whether these semantic contents are in fact propositions; in particular, they may very well be events (or 'eventualities').

Contrary to what is sometimes believed, semantic rhetorical relations are not always relations between events (or 'eventualities'). Consider the following example, where A and B discuss the use of remote controls:

(9) A: You keep losing them
B: That's because they don't have a fixed location

The 'event' in the second utterance (*having a fixed location*) does not cause the *losing* event in the first utterance; on the contrary, the second utterance says that the fact that *no* having-a-fixed-location event occurs is the cause of the *losing*. Saying that a certain type of event does *not* occur is not describing any event, but expressing a *proposition* (about that type of event). This means that the causal connection between the two utterances is not between two events, but between the *proposition* made in the second utterance and the event in the first utterance.[6]

Rhetorical relations between dialogue utterances do not necessarily relate the *semantic contents* of the dialogue acts that they contain, but may also relate the *dialogue acts* as such, taking both their semantic contents and their communicative functions into account. The following examples[7] illustrate this:

(10) A1: Where would you physically position the buttons?
A2: I think that has some impact on many things

(11) B1: I'm afraid we have no time for that.
B2: We're supposed to finish this before twelve.

Utterance A2 in (10) encodes an Inform act which has a *Motivation* relation to the Question act encoded in A1; it tells the addressees what motivated A to ask the question A1 with this particular semantic content. In (11) utterance B2 encodes an Inform act which has an *Explanation* relation to the Decline Request act in B1.

[6]It could in fact be argued that the first utterance also contains a proposition, rather than describing an event.

[7]From the AMI meeting corpus - ES2002a.

5 Scope and distance

While a feedback dependence relation can target an utterance, a functional segment, a dialogue act, a turn unit, or a group of those, functional dependence and rhetorical relations are grounded in meaning and follow more restricted patterns of linking. We investigated the linking patterns of the different types of relations for two corpora of annotated dialogues, the AMI meeting corpus and a French two-party route explanation dialogues collected at the University in Toulouse[8].

For analysing these patterns it is helpful to look at the *scope* and *distance* covered by a relation. We define scope as follows:

(12) *the scope of a discourse relation is the number of functional segments (the 'target') that a given segment (the 'source') is related to.*

Calculation of the distance between related functional segments in dialogue is not a trivial task and deserves some discussion. The distance between two segment can be calculated *textually*, e.g. as the number of intervening constituents between a pair of constituents under consideration, or as the number of intervening tokens; and *topologically*, as the length of the shortest path in a graph representation (e.g. a SDRS, see Afantenos and Asher, 2010). Since in this study we did not construct any tree or graph representations for the various kinds of relations we distinguished, we considered the textual calculation of distance between related segments. In dialogue, the most plausible unit for measuring distance is the functional segment, but simple count of intervening functional segments is not possible, because of the following complications:

- Spontaneous speech includes self-corrections, retractions and restarts that have a communicative function of their own and are considered as functional segments. Speech errors and flaws like reparanda (segment parts that are retracted or corrected by the same speaker) do not have any communicative function on their own;
- Functional segments may be discontinuous and may be interrupted by more substantial segments than repairs and restarts, e.g. 'Because twenty five Euros for a remote... *how much is that locally in pounds?* is too much

[8]For more information see (Muller and Prévot, 2003) and http://crdo.fr/crdo000742

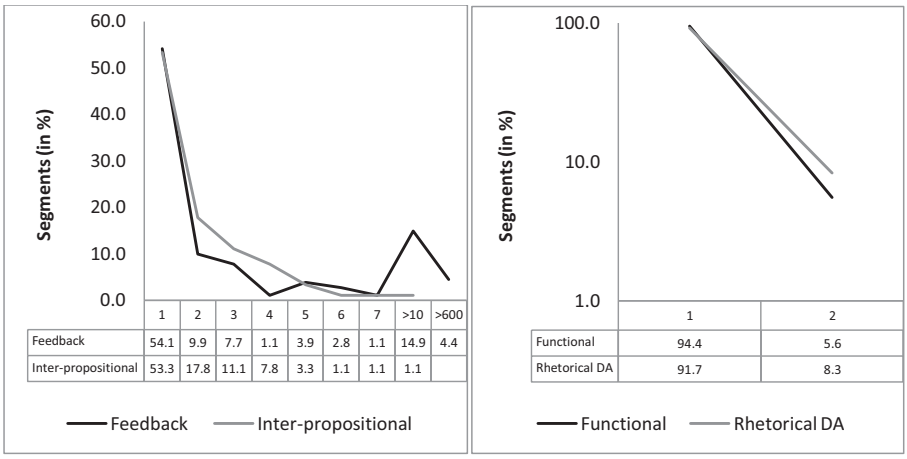


Figure 1: Scope of feedback dependence, functional dependence and rhetorical relations in the AMI data.

money to buy an extra remote or a replacement remote';[9]

- Functional segments may be spread over more than one turn, e.g. A: Well we can chat away for ... um... for five minutes or so I think at... *B: Mm-hmm* ... at most;[10]
- Functional segments may overlap, e.g. U: What time is the first train to the airport on Sunday? S: *The first train to the airport on Sunday* is at 6.17, where the part in italics forms part of an answer to U's question, but also has a feedback function, displaying what S has heard;
- In multi-party interaction multiple participants may react to the speaker's previous contribution and may do this simultaneously, with some overlap or consecutively, e.g.

(13) A: Do you have anything to add?
B: *No*
C: *No*

These dialogue-specific phenomena should be taken into account while calculating the distance between related functional segments. All segments were ordered by their start time. Given two non-overlapping segments A and B, with begin(B) = end(A) (i.e. B starts after A has ended) a segment C is counted as intervening between A and B if and only if C starts later than A and ends before B. (In that case, C must contain some material that occurs after A has ended and before B has started) We thus define:

(14) *A segment C intervenes between the segments A and B iff begin(C) > begin (A) and end(C) < end(B).*

Moreover, if C and D are two intervening segments with the same begin- and end points, with the same communicative function(s) and with identical semantic contents (but contributed by different speakers), (cf. (13)), then they are counted only once; and if an intervening functional segment E is a sub-segment of a larger intervening segment K produced by the same speaker, the only the larger segment is counted.

If A and B are overlapping or consecutive segments, i.e. begin(B) <= end(A), we stipulate their distance to be zero. Hence we use the following definition of distance:

(15) *The distance between two non-overlapping segments A and B, with begin(B) > end(A) equals the number of intervening functional segments* minus *the number of co-occurring intervening functional segments with identical wording and interpretation (produced by different speakers)* minus *the number of sub-segments of intervening functional segments produced by the same speaker.*

Moreover, in order to deal with the complications mentioned above, we removed all reparanda from the data, e.g. 'This is the kick-off meeting for our our project' became 'This is the kick-off meeting for our project'; and we merged functional segments that were spread over multiple turn units.

Figure 1 shows the scope and Figure 2 the distance involved in functional and feedback dependence relations, and for inter-propositional relations and rhetorical relations between dialogue acts, as found in the AMI corpus. Our analyses show that different relations exhibit different patterns. A functional dependence relation normally has a narrow scope (1-2 functional segments), and units related by this type of relation tend to be close to each other in discourse, except in the case

[9] From the AMI meeting corpus - ES2002a.
[10] From the AMI meeting corpus - ES2002a.

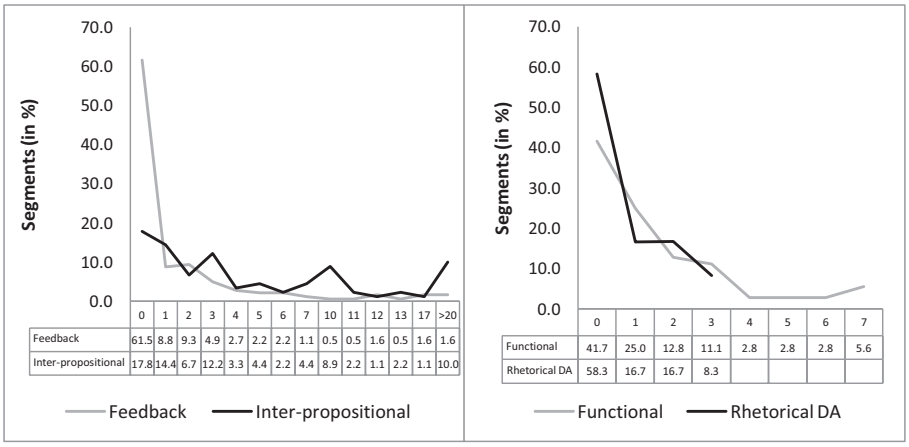


Figure 2: Distance of feedback dependence, functional dependence and rhetorical relations in AMI data.

of discourse pop-up units. Feedback dependence relations as a rule have either very narrow or local very wide scope; long-distance attachments are rare. Feedback acts can target all types of dialogue units that we have defined: other dialogue acts, turn units, functional segments, as well as groups of those. Rhetorical inter-propositional relations often have narrow scope but the related segments may be some distance away from each other. Rhetorical relations between dialogue acts are characterized by (as a rule) narrow scope and short distance, but some rhetorical relations (like *Recap, Conclude*) link a dialogue act or a dialogue act group to one or more dialogue act groups, having a wide scope.

Four types of attachment in terms of distance and scope can be distinguished for the way in which discourse relations connect a source segment to other units in dialogue:

1. Last segment: A relation links the source segment to the previous functional segment: scope is 1 and distance is 0.
2. Local attachment: The source segment is related to several previous segments. The scope of each relation is 1; at least one of the relations has a distance 0, and at least one has distance greater than 0. For example, the next step of a narration introduces a contrast with the preceding segment, while elaborating an earlier segment.
3. Local wide scope attachment: The relation targets a group of segments. The scope is larger than 1, the distance is 0. This is common with relations such as *Recap*, *Summarize*, *Conclude*.
4. Discourse pop-up: The source segment is related to an earlier functional segment or to a group of functional segments.

Attachments of type 1 occur frequently (29.8% of all attachments in the AMI corpus). For example[11]:

(16) D1: Cost like production cost is twelve fifty or retail like on the shelf?
B1: Our sale anyway
B2: Because its probably up to the retailer to sell it for whatever price they want

Segment B1 is an Answer to the Choice Question in D1, and segment B2 provides a Justification for the Answer in B1.

Attachments of type 2 are more complicated. Such attachments are frequently observed in the AMI data and account for 41.5% of all attachments. For example[12]:

(17) D1: Now remote controls are better
D2: but actually still kind of like a massive junky thing [Contrast:D1]
B1: Still feels primitive [Elaboration:D2;Contrast:D1]

The fact that the related segments are produced by different speakers has the consequence that they exhibit not only rhetorical but also feedback dependence relations by implication, e.g. the expression of Agreement in B1 implies positive feedback on understanding D2.

Local wide scope attachment is frequently observed for feedback dependence relations. Very often feedback is provided not to a single functional segment but to the discourse unit that is concerned with one of the dialogue sub-tasks or

[11] From the AMI meeting corpus - ES2002a.
[12] From the AMI meeting corpus - ES2002a.

topics. This occurs frequently in multi-party dialogues (19.2% in the AMI meetings). Both positive and negative feedback are observed to sometimes have local wide scope attachment. For example[13]:

(18) B1: We're gonna be selling this remote control for twenty five euro
B2: and we're aiming to make fifty million euro [Narration:B1]
B3: so we're gonna be selling this on an international scale [Elaborate:B1&B2]
B4: and we don't want it to cost more than twelve fifty euros [Narration:B3]
D1: Okay [PositiveFB:B1-B4]
B5: So fifty percent of the selling price [Conclude:B3&B4]
D2: Can we go over that again [NegativeFB:B1-B5]

Feedback elicitation acts may also have local wide scope. Examples are: *Does anybody have anything to add to the finance issue at all?, Anybody anything to add here?* and *Any thoughts about this?* This often happens when moving from one topic to another.

Muller and Prévot (2003) have shown that in French route explanation dialogues, *voilà (that's it)* is a marker of closure, thus being some kind of wide-scope feedback (type 3) preparing a discourse pop-up of type 4. For example[14]:

(19) B1: So I guess that's it
D1: Great
B2: The meeting is over
B3: Whoohoo

Rhetorical relations may also have scope over larger units, like dialogue act groups. Concerning rhetorical relations like *conclude*, *recap*, and *summary*, it may be noticed that a conclusion, a recap, or a summary is often not expressed by a single dialogue act but by a group of them. In such a case we need to allow for rhetorical relations between dialogue act groups. For example, in AMI meeting ES2002b after a discussion stretching over some 150 segments about the functionality to be included in a remote control, the participants came to a conclusion proposed in D1-D9 and acknowledged in B5-C2:

(20) D1: Well we want this to be a product that offers simple and all the sort of more tricky features
D2: but we want them to be in another area
B2: Think what we absolutely have to have and what would be nice
B3: To recap you've got volume and channel control
C1: There's on and off
B4: Volume and channel and skip to certain channels with the numbers
D3: rarely used functions may be in a little area but covered up
D4: things like channel and volume are used all the time
D5: We just have them right out on top
D6: so we need to think about having three or more groupings of controls
D7: like one which are the habitual ones that should be right within your natural grip
D8: ones that with available features
D9: And then others with concealed
B5: Okay
B6: Any of you anything to add to that at all?
A1: No
C2: No

Discourse pop-up attachments, finally, are especially observed for rhetorical relations. Consider the following example[15]:

(21) G13: hop hop hop Esquirol tu continues tout droit (hop hop hop Esquirol continue straight)
G14: y'a le Classico (there is the Classico)
R15: euh (uh)
G16: t'as pas l'air branché trop bars (you do not seem to be into bars)
R17: euh non (uh no)
R18: mais je connais pas très bien Toulouse (but I don't know Toulouse very well)
G19: ah ouais d'accord (ah yeah ok)
G20: donc Les Carmes tu vois ou c'est? (so Les Carmes, you know where it is?)
F21: oui (yes)
G22: bon ben voilà. (well that's it)
G23: donc là tu continues sur sur cette rue (so there you continue on this street)
G24: et tu arrives aux Carmes (and you get to Les Carmes)

Segment G22 concludes and closes discourse unit [G14-G21], and a Continuation/Narration relates G13 to G23.

Many discourse markers, which have been studied for their semantic contribution and for their role in dialogue structuring, are good indicators for various kinds of discourse attachment. Most connectives (*then, but, therefore*) connect functional segments with attachment of type 1 or 2. Enumerative markers such as (*First, Then, Finally*) can introduce macro-structures resulting in both long-distance and local wide-scope attachment, since usually the entire discourse unit that contains these markers is rhetorically related to another discourse unit.

The findings discussed here are summarized in Figure 3, which shows an ISO-style metamodel (cf. Bunt et al., 2010) containing the various kinds of units in dialogue and the possible relations between them.

[13] From the AMI meeting corpus - ES2002a.
[14] From The AMI meeting corpus - ES2002b.
[15] From the French route navigation corpus.

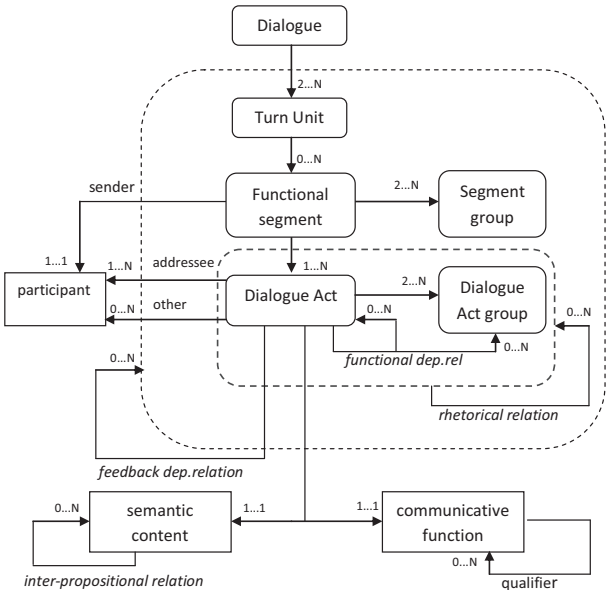


Figure 3: Metamodel of dialogue units and relations between them.

6 Conclusions and future research

In studying the occurrence of discourse relations in dialogue, we have observed at least four types of relations: rhetorical relations between dialogue acts or between their semantic contents (inter-propositional rhetorical relations); feedback dependence relations; and functional dependence relations between dialogue acts. Some of these relations may also involve larger units or groups of those, and we have seen that the various kinds of relation show significant differences in scope and distance of attachment.

In future studies of the properties of discourse relations in dialogue, such as their scope and distance, it may be useful to distinguish between the semantic and presentational dimensions of rhetorical relations. It may also be noted that the metamodel in Figure 3 has been designed in such a way that it is an extension of the metamodel for dialogue act annotation used in ISO DIS 24617-2 (see Bunt et al., 2010). According to the dialogue act-theoretical framework that we have used, the semantic content of a dialogue act is typically an eventuality or a proposition. Adding this distinction to the metamodel would open the way for connecting with the metamodels used in ISO projects concerned with the semantic annotation of time and events, space, and semantic roles, which could be very helpful for clarifying the relations between the semantic phenomena targeted in these projects.

References

Allwood, J. 1992. *On dialogue cohesion*. Göthenburg Papers in Theoretical Linguistics 65. Gothenburg University, Department of Linguistics.

Afantenos S.D., and Asher N. 2010. *Testing SDRT's right frontier* Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, pp. 1-9.

Asher, N., and Lascarides, A. 2003. *Logics of Conversation*. Cambridge University Press, UK.

Bunt, H. 2000. *Dynamic Interpretation and Dialogue Theory*. In M.M. Taylor, D.G. Bouwhuis and F. Neel (eds.) The Structure of Multimodal Dialogue, Vol 2., Amsterdam: John Benjamins, pp. 139-166.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang,A., Hasida, K., Petukhova, V., Popescu-Belis, A., Soria, C., and Traum, D. 2010. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO document ISO/TC 37/SC 4 N442 rev 05*. ISO Central Secretariat, Geneva.

Carlson, L., Marcu, D., and Okurowski, M.E. 2001. *Building a discourse-tagged corpus in the framework of rhetorical structure theory*. Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, pp. 1-10.

Clark, H. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

van Dijk, T. A. 1979. *Pragmatic connectives*. Journal of Pragmatics, 3: 447-456.

Geertzen, J., Petukhova, V., and Bunt, H. 2007. *A Multidimensional Approach to Utterance Segmentation and Dialogue Act Classification*. Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, pp. 140-149.

Grosz, B., and Sidner,C. 1986. *Attention, intentions, and the structure of discourse.* Computational Linguistics, 12(3): 175-204.

Hobbs, J. 1979. *Coherence and Coreference*. Cognitive Science, 3:67-90.

Hobbs, J. 1985. *On the Coherence and Structure of Discourse*. Research Report 85-37, CSLI, Stanford.

Hovy, E.H. 1990. Approaches to the Planning of Coherent Text. In Swartout, C.L. and Mann, W.C.(eds.) Natural Language in Artificial Intelligence and Computational Linguistics, Kluwer, Boston, pp. 83-102.

Hovy, E.H. 1995. *The multifunctionality of discourse markers*. Proceedings of the Workshop on Discourse Markers, Egmond-aan-Zee, The Netherlands.

Hovy, E., and Maier, E. 1995. *Parsimonious of profligate: how many and which discourse structure relations?* Unpublished manuscript.

Larsson, S , and Traum, D. 2000. *Information state and dialogue management in the Trindi dialogue move engine toolkit*. Natural Language Engineering, 6(3-4): 323-340.

Louwerse, M., and Mitchell, H. 2003. *Toward a taxonomy of a set of discourse markers in dialogue: A theoretical and computational linguistic account*. Discourse Processes, 35(3): 243-281.

Mann, W. and Thompson, S. 1988. *Rhetorical structure theory: toward a functional theory of text organisation*. The MIT Press, Cambridge, MA.

Miltsakaki, E., Prasad, R., Joshi,A., and Webber, B. 2004. *The Penn Discourse Treebank*. In Proceedings of International Conference on Language Resources and Evaluation.

Moore, J., and M. Pollack. 1992. *A problem for RST: The need for multi-level discourse analysis*. Computational Linguistics, 18:537-544.

Muller, P., and Prévot, L. 2003. *An empirical study of acknowledgement structures*. Proceedings of the 7th Workshop on Semantics and Pragmatics of Dialogue, Saarbrucken.

Nakatani, C., and Traum, D. 1999. *Draft: Discourse structure coding manual. Version 1.0. Technical Report UMIACS-TR-99-03*. University of Maryland.

Petukhova, V., and Bunt, H. 2009. *Towards a Multidimensional Semantics of Discourse Markers in Spoken Dialogue*. In Bunt, H., Petukhova, V., and Wubben S. (eds.) Proceedings of the Eighth International Workshop on Computational Semantics (IWCS),Tilburg, pp. 157-168.

Poesio, M., and Traum, D. 1998. *Towards an axiomatization of dialogue acts*. Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues, pp. 207-222.

Polanyi, L. 1988. *A formal model of the structure of discourse*. Journal of Pragmatics, 12: 601-638.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo,L., Joshi,A., and Webber, B. 2008. *The Penn Discourse Treebank 2.0.* In Proceedings of the 6th International Conference on Language Resources and Evaluation.

Reese B., Denis P., Asher, N., Baldridge. J., and Hunter, J. 2007. *Reference Manual for the Analysis and Annotation of Rhetorical Structure (Version 1.0)*. University of Texas

Sanders, T., Spooren, W., and Noordman, L. 1992. *Toward a taxonomy of Coherence relations*. Discourse Processes, Vol. 15, pp. 1-35.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: University Press.

Stede, M. 2008. *Disambiguating Rhetorical Structure*. Research on Language and Computation, 6(3-4): 311-332.

Traum, D. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. PhD Thesis, Department of Computer Science, University of Rochester.

Wolf, F., and Gibson, E. 2005. *Representing discourse coherence: A corpus-based study*. Computational Linguistics, 31(2): 249-287.

A deep Ontology for Named Entities

Gil Francopoulo
Tagmatica
126 rue de Picpus
75012 Paris France
gil.francopoulo@tagmatica.com

François Demay
(Independent Consulting)
26, rue de Rochefort
78120 Clairefontaine France
francois@edemay.com

Abstract

The number of documents published every day on the web becomes huge. Manual construction of metadata for each page being hard and unreliable, automatic semantic annotation seems to be the unique practical solution to meet these increasing needs for information extraction and retrieval. With this respect, ontology design plays a key role: a common ontology permits a good interoperability between software components at the level of semantic annotations, but also allows the final user to easily interpret the result of the semantic extraction. In this paper, we present an operational, industrial and extended ontology for newspapers and blogs processing.

1 Introduction

Ontologies for named entities (NE) are designed to meet increasing needs for NE types. These ontologies originated from the set defined by MUC [Grishman 1996] with 7 categories (people, organization, location, time, date, money, percentage expressions). The number of types was increased to 93 (29 types and 64 subtypes) by BBN [Brunstein 2002]. The types were extended in several steps by Sekine to reach 150 and then 200 NE types [Sekine 2002, 2004][1]. For the history of NE typing, see [Ehrmann 2008].

2 Objectives

Just dealing with newspapers, newswires and blogs, the number of documents published every day on the web becomes huge. Only a subset of these contents is annotated with reliable metadata, and, of course, except for certain known sources, frequently, we don't know whether the metadata is correct or wrong[2]. Another problem is that the metadata tags are heterogeneous and, therefore difficult to compare. The most reliable material still remains the text content.

Our aim is to filter, dispatch and classify documents for professional watchers. Named entities are very good clues for this purpose. From the point of view of the user, very simple algorithms may be implemented: one of them may be for instance to consider that if, in a document is met the name of a soccer player or a soccer club, this document deals with soccer.

3 Situation

The problem we faced was two-fold. First, the 200 Sekine's types were not precise enough in the micro-domain where the system has a rich net of information and where the requirements from our professional users are rather high. Secondly, it was not possible to adopt an ontology not suited for NE like the hierarchy of types of WordNet. As discribed in section-10 (ambiguities), we need some specific intermediate types in order to retain a certain level of indetermination. For instance, a splitting at a higher level between abstract and concrete objects is not useful. On the contrary, a node merging the geographical and the political aspect of a city is useful because this node provides a geopolitical type when the geographical and political aspect cannot be distinguished.

[1] See also the beta English V7 released when we wrote this article on http://nlp.cs.nyu.edu/ene

[2] For instance, very basic tags like language names are sometimes wrong. Typically, a page is written in one language, the content is translated and the original metadata is not modified accordingly.

Another problem to solve was more on the human side than the scientific one. In a professionally system, some way or another, the type is presented to the user together with the ontology. That means that the user must understand the meaning of the type. The situation is different from the one in the 90's where an Academic player could develop a system in a laboratory and then propose the resulting system to professional users. Now watchers have ontologies, best practices habits, professional associations like IPTC[3] that recommend lists of codes. This new context has an impact on the methodology: the current users are rather mature and want to be associated with the design.

This is not to say that the situation is idyllic. Some users find that the professional ontologies cover the majority of their needs but some others think that the recommended lists and structures are too complex. We tried to accommodate with this situation.

Our strategy was to build the ontology by hand from existing lists, with a co-operation between experts of encyclopedia writing, NLP practitioners and professional watchers. The role of an expert in encyclopedia writing is to propose an organization of types. The role of a NLP practitioner is to test against real texts to verify how the type matches against named entities. The role of the professional watcher is to verify whether the result is understandable.

4 Method for building the ontology

The main burden is to propose an organization of types.

The first step was to split the domain according to Sekine's high level types. This is not to say that we did not modify afterwards these assumptions, but we took these types as a starting point. These modifications have been applied based on experiments on unknown named entities as explained in the following section "The ontology".

Starting from a type, a series of steps were tempted.

Step#1: is there a sub-tree which may be taken from a recommended professional list?

Step#2: is there a sub-tree in an encyclopedia? Usually, a kind of classification of the articles has been implemented in order to help the managers / editors master a certain balance between subjects or a relative completeness. These classifications were always some kind of topic tree where each level of the hierarchy could mix different nature of information: nature of the item being written about, place, time etc. To have a much better way to master the content, this type of classification can be replaced and complemented by a more systematic categorization scheme where each article would be located in a multidimensional space: each coordinate of this space would correspond to the answer (for the item) to the W questions : what / who is it?, where is it located in a subject / knowledge cartography?, where is it located in the (present or a previous) geopolitical area?, when is the item taking place? When this kind of categorization is implanted it becomes very easy to make a thorough analysis of the named entities, for proper names, strictly speaking, at least.

Step#3: Another way, when no such possibility exists as explained above, one can use the first sentence of the article of the encyclopedia: all the most relevant features defining the item are there, and one can retrieve a lot of metadata leading usually to the named entities characterization.

Step#4: Is there is any fact box (infobox in Wikipedia's terminology) that may also be used?

Step#5: Last, when categories have been given to the article, some of these may also be used. In Wikipedias, this has been done. Unfortunately, there has been usually no consistence in the rules for the categorization (and their implementation) even for articles in one and the same language and much more less consistency between categories in different language contents.

One of the author of the current work has been able to implement the systematic categorization (as described above with multidimensional localization) to a certain number of big encyclopedic contents in French, English and German for the Encyclopedia Universalis (15 years, since 1965), Larousse (20 years), Encarta (3 years) and then for a Chinese-French lexicon. This experience of 40 years has been very useful in building by hand the current ontology.

[3] www.iptc.org

5 Design principles

The Tagmatica's deep ontology for named entities is designed mainly for information extraction from newspapers, newswires and blogs. The type names are labeled in English and the meaning of each type is the definition in English but we took great care to respect a language neutrality because the NE extraction is not restricted to English. Currently, the same ontology is used for three languages: French, English and Spanish. We didn't have the need to create any specific sub-tree for a given language. Three domains are addressed on a fine grain basis: politics, economics and sports. These domains are rather general and universal. In other terms, we do not address technical domains like genomics or mechanics.

We distinguish the notion of type from the notion of role. A type (and sub-type) is considered as a rigid subdivision, in the sense that this is the reason why an entity is known. For instance, "Jacques Chirac" is considered as a politician, with the convention that we are dealing with the famous human being and not with an unknown person whose name is "Jacques Chirac", i.e. a homonym. This labeling is considered as type labeling because this is the reason why we know his name. We don't consider the fact that when he was a child (or a baby) he was not a politician. On the contrary, his role as president is not considered as a type. This information is managed at the instance level as a function name together with starting and ending dates.

6 The ontology

Our ontology is designed having in mind the named entity recognition (NER). This process faces two different situations: i) the name (or fragments of the name) is (or are) already recorded in the system and successfully recognized, ii) the name is unknown, which means that the immediate context must be interpreted in order to determine a type for the name.

Let's take two examples:

1) "Messi is supported in this by FC Barcelona, ...". With the convention that "Lionel Messi" is recorded in the database, the NER recognizes and determines that "Messi" is the name of a soccer player. Thus, a great level of detail can be determined, including the variants like "Leo Messi" (his usual name), "Lionel Messi" or "Lionel Andrés Messi" (his official full name) together with an URI to a Wikipedia entry for additional documentation.

2) In another situation, with the convention that "Marcel Dujardin" is an unknown person, the sentence "Marcel Dujardin drove too fast ..." cannot give such a level of detail. But, provided that "Marcel" is recognized as a male given name and the first letter in "Dujardin" is an uppercase one, the NER is able to determine that "Marcel Dujardin" is a pair of words combining a given name and a family name, i.e. a person name. But, aside from the sexual genre, we cannot determine any information about the usual activity of this person.

Based on the requirement that we must deal with both known and unknown names, we decided to design a deep ontology for types dedicated to known names and a level-1 types for unknown names. We structured the first level to determine a type in case of an unknown name. We may add, that in some cases, it is possible to determine a slightly more precise typing than just the first level, but the precision does not go very far. For instance, a flight identification code is recognized as a pair of specific uppercase letters (to be taken in a pick-list) and three digits. In other terms, the system is always able to determine a main type (the first level) and optionally, in the most favorable situation, a sub-type can be determined.

We defined and developed an ontology of 995 types based on Sekine's hierarchy, IPTC event types [EventsML-G2], geonames[4] and previous works in encyclopedia structuring.

7 First level types

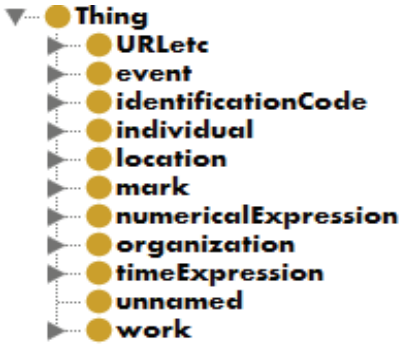
The level-1 has 11 types, as follows:

- **URLetc**, for filenames and URL,
- **event**, for event names like "Tour de France". This sub-tree is taken from the IPTC's registry for the types of events. This thematic classification is rather usual in the domain of newswires and professional watchers.
- **identificationCode** for all alphanumerical codes like flight number (e.g. AF 447) or ISBNs.
- **individual**, for an individual person. The person may be living (or has lived) or imaginary.

[4] www.geonames.org

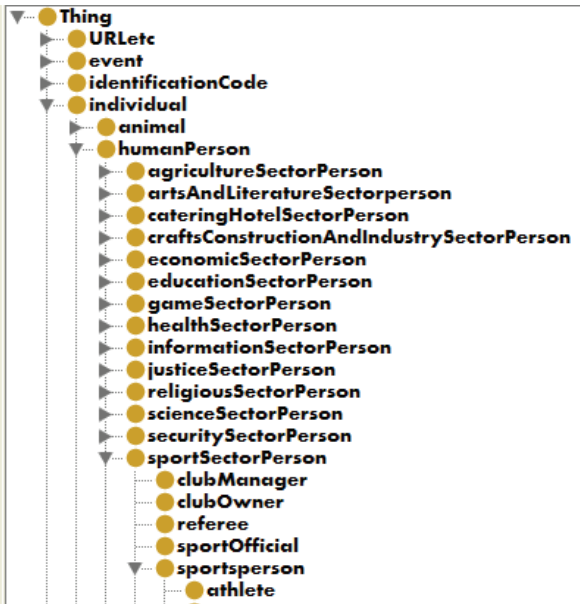
It is in general the name of a human being but the type may be used also for a pet or a plant. This type is not to be used for a group of people, see the organization item for this purpose.
- **location**, for geopolitical, geological and geographical entities. Continents and planets are also covered by this type.
- **mark**, for mark names like commercial trademarks, formats and protocols. Let's note that when the name is both a mark and the name of an organization, we adopt the convention that the name should be labeled as an organization.
- **numericalExpression**, for all forms containing a number with or without a unit. Examples are measures and percentages. This type is equivalent to NUMEX in MUC.
- **organization**, for a company, institution or a group of people.
- **timeExpression**, for a reference of time like dates, times, combinations of date and time. This type is equivalent to TIMEX in MUC.
- **unnamed**, for all the common nouns that are not in the other types and that are used as head of a noun phrase in the corpus. The main objective being to label named entities, this type cannot be used to directly mark a named entity. By means of the coreference, such a noun may be used to indirectly designate a named entity. For instance, in the text "Chirac ... The president ...", the function name "president" will be marked as unnamed and the coreference resolution module will link "Chirac" and "president"[5].
- **work**, for names of human works like movies, books, sculptures, songs etc.

To summarize, the first level is as follows in Protégé[6]:



8 One example

It is not possible to present in detail all sub-trees, so we will present only one example with a five-level depth for athlete:



9 Other levels

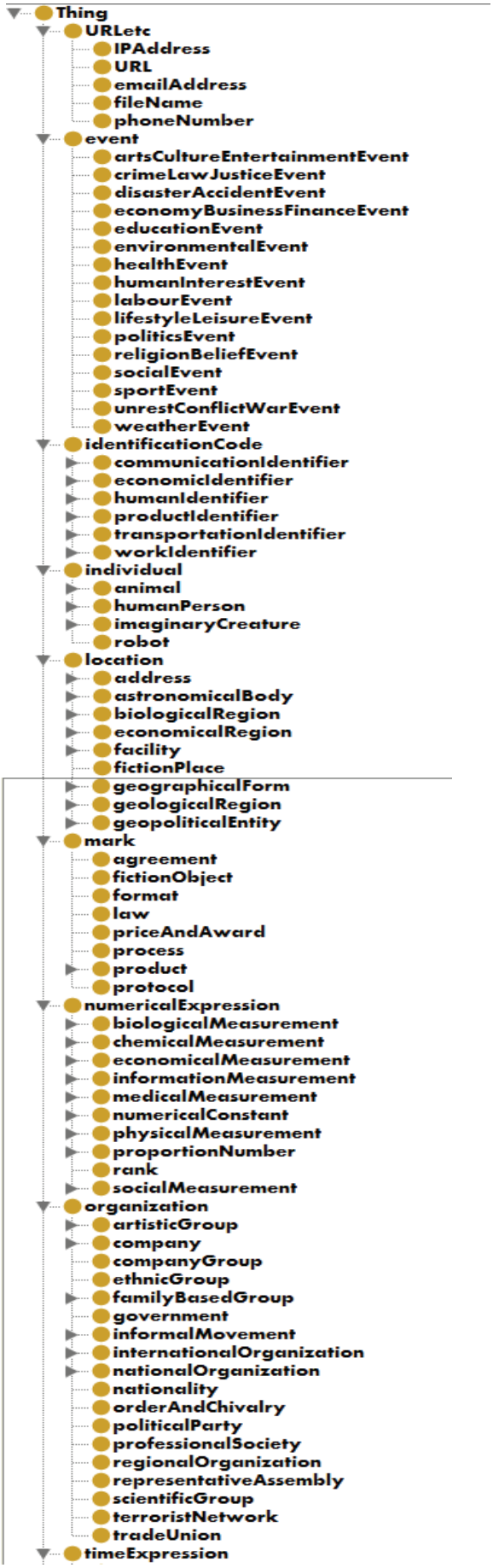
Depending on the types, the deepness of the hierarchy is between two and five. The ontology being rather large, we advice the interested reader to directly download the OWL file and use Protégé to browse through the different levels. The ontology is freely available on the Tagmatica's website[7].

The level-2 is as follows:

[5] As the notion of named entity has been extended from proper name to time expression in the 90's, we extend the notion of named entity to "non-proper" references to a named entity.

[6] http://protege.stanford.edu

[7] http://tagmatica.fr/doc/ontology.owl



10 Ambiguities

As mentioned in [Sekine 2002], "Japan" is normally used in a geographical sense, but sometimes it refers to the government of Japan (organization), as in "Japan announced a tax cut". This problem is a rather general problem that does not concern only the country names but concerns also all cities and villages. The NER is not able to distinguish the two senses on a reliable manner. So, we consider "Japan" as geographical and political entity (the GPE in ACE definition).

We adopt the same strategy for entities like airports that may be considered as geographical or as organizational entities. All these items are recorded under the node "facility". Instead of focusing on where and what the entity is, we prefer the usage aspect: "what is it for?".

11 The instances

The ontology is a hierarchy of types which is built by hand and is rather stable. On the contrary, the proper names (i.e. the instances), are automatically collected from different Wikipedia dumps[8] and may change frequently because new proper names appeared every day. For this purpose, a series of filters have been coded in Java in order to associate the types with the field names of Wikipedia's infoboxes. Most of the time, an instance is associated with only one type, but they are some exceptions like a famous judoka who is also deputy. We collect three Wikipedias (in French, English and Spanish), and from each file, we extract a selection of proper names. Then, the three results are merged together and finally merged with the current database. At present, we collect these names when needed but, in the near future, we plan to collect fresh data every week-end, systematically in order to be synchronous with new names. Let's add that a certain number of locations have been extracted from the

[8] http://download.wikipedia.org/backup-index.html

Geonames site but this work is not finished. At the moment, the total number of instances is 200 000.

This requirement to stick to the current state of the art of all fresh data published on the web has a certain number of constraints concerning the choice of the sources for updating the instances. Let's recall that our main application is newspapers processing. Wikipedia dumps are updated every four days (in average and when everything goes fine) without a precise date for each dump. Our instances will be updated every week-end, so the data are synchronous enough[9]. On the contrary, we cannot download DBpedia (see http://dbpedia.org) and the web of data which is computed from DBpedia (see http://linkeddata.org) because sources like DBpedia are updated every six months. This is for us a too long time. For the same reason, this requirement prevents us to use gazetteers included within frameworks like Gate (see http://gate.ac.uk). The update frequency is too low.

12 NE extraction

The NE extraction is not a stand-alone software module. It's a component of an hybrid industrial parsing scheme combining Hidden Markov Model implementations [Bikel 1997] and hand-written rules. The system been described elsewhere [Francopoulo 2008], we are not going to present the modules in detail. We may just add that the main modules are based on active learning techniques and that the whole system is a pipeline of modules for language detection, error recovering, chunking, syntactic parsing, coreference resolution and quotation extraction in a robust manner. The ontology of types and the instances are shared by the three processed languages. Each language is described in a specific lexicon called TagDico that conforms to the ISO standard ISO-24613 for NLP lexicons: LMF (for Lexical Markup Framework) [Francopoulo 2006].

13 Relation with standards

The NE extraction is consistent with the ISO Preliminary Work Item for the representation of named entities: ISO 24617-3 where the entities are annotated in a stand-off scheme in the spirit of the Linguistic Annotation Framework (LAF, i.e. ISO-24612) [Ide 2004].

[9] If we discover that it is not the case, we could refresh every day: the process is fully automatic.

With this respect, the labeling of NE is more powerful than inline annotation for difficult annotations in which the elements are not contiguous like "Bill and Hillary Clinton" and where the NER must detect two named entities with a distribution of the family name to the two given names. More traditional systems like BBN Named Entity Annotation (see www.anc.org/annotations.html#bbnne, for instance) cannot deal with such annotations because they are inline based. It should be noted that most systems do not deal with these problems (see [Erhmann 2008] for a discussion). The objective being to build an index, the two named entities must be recognized by the system.

Another difficult problem arises when one entity of a certain type is a sub-part of another entity with a different type. For instance, in: "the city of Michelin ..." where "city of Michelin" is a geopolitical entity (as a city) but where "Michelin" is the name of an organization. Let's note, that if an inline annotation scheme is used but with the option of embedding different levels of annotation, the annotation is possible, on the contrary of the first example where there is no way to address the problem. Again, the objective being to build an index, the two named entities must be recognized by the system.

Concerning XML serialization, the physical file is coded in OWL as defined by W3C at www.w3.org/2004/OWL.

14 Evaluation

There is no quantitative evaluation. Evaluation is important, but we have no budget for this task and this is not the right period. Our users make comments and the system is modified almost every day.

15 Conclusion

Following rather practical lines of action and after some long talks and negotiations based on an extended experience of ontology structuring, we created a deep ontology of types for named entities representation and automatic recognition with a fine set of interoperable semantic annotations.

The domains we currently address being politics, economics and sports, we targeted rather general domains. We don't claim that our

ontology could be easily extended to include deep technical domains like genomics or mechanics. At first view, these domains require separate and specific ontologies. But what is possible, is to extend selected sub-trees to cover more deeply a specific application domain. In the past, we already extended successfully some sub-trees for specific needs like airline, soccer or athletics domains, without any problem.

Our ontology begins to be effectively used and some sub-parts may need to be tuned or extended in the near future based on the user' feedbacks. We welcome all comments and useful suggestions.

Acknowledgements

The work presented here was partially funded by the System@tic competitiveness cluster http://www.systematic-paris-region.org within the Scribo project, see www.scribo.ws.

Reference

Bikel D., Miller S., Schwartz R., Weischedel R. 1997 "Nymble: a High-Performance Learning Name Finder" ANLP-1997.

Brunstein Ada 2002 "Annotation guidelines for answer types", BBN technologies report at www.ldc.upenn.edu/Catalog/docs/LDC2005T33 (on Dec 2010)

Ehrmann M. 2008 "Les entités nommées, de la linguistique au TAL", PhD thesis, Univ. Paris 7.

EventsML-G2 International Press Telecommunications Council (IPTC) www.iptc.org

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF), LREC-2006, Genoa.

Francopoulo G. 2008 TagParser: well on the way to ISO-TC37 conformance. ICGL-2008, Hong Kong.

Grishman R., Sundheim B 1996 "Message Understanding Conference - 6: A Brief History", COLING-96.

Ide N., Romary L. 2004 International Standard for a Linguistic Annotation Framework, Journal of Natural Language Engineering, 10:3-4, 211-225.

Sekine S., Sudo K., Nobata C. 2002 Extended Named Entity Hierarchy, LREC-2002, Las Palmas.

Sekine S., Nobata C. 2004 Definition, dictionaries and tagger for Extended Named Entity Hierarchy, LREC-2004, Lisbon.

Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems

Thierry Declerck
DFKI GmbH, LT-Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbruecken
declerck@dfki.de

Piroska Lendvai
Research Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33., H-1068 Budapest
piroska@nytud.hu

Tobias Wunner
DERI, NLP unit
NUIG
Newcastle Rd, IE-Galway
tobias.wunner@deri.org

Abstract

We investigate the benefits that can result from the formal representation of linguistic and semantic features of natural language expressions that are used as terms in labels of knowledge representation systems, like taxonomies and ontologies. We show that such a representation can support Human Language Technologies and Semantic Web applications, especially in the context of ontology-based information extraction, since it gives a basis for specifying mapping strategies between the restricted natural language used in taxonomies and ontologies and the unrestricted language used in documents processed by information extraction or semantic annotation tools.

1 Ontology-based Information Extraction

In the last decade, we have been witnessing changes in the field of Information Extraction (IE) due to the emergence of a significant amount of semantic resources available in the form of taxonomies and ontologies. These Knowledge Representation (KR) systems have been gradually replacing the predefined templates, which were formerly used for specifying IE applications, and are now often building the background against which texts are processed in order to extract relevant information for specific applications. In those cases, we speak of Ontology-based Information Extraction (OBIE)[1].

In the Description Logic (DL) approach, KR systems are viewed as consisting of two components, the T-Box (Terminological axioms) and the A-Box (Assertion axioms).[2] We adopt here this terminology (*T-Box*, *A-Box*), even if not all the KR systems we are dealing with are modeled using the DL representation language, and in fact we are dealing in this short paper only with examples taken from a complex taxonomy modeled in XML.

A main issue for OBIE tasks is to establish an accurate mapping between the classes and properties described in a T-Box and the natural language expressions occurring in unstructured textual documents. Fortunately, most KR systems come equipped with a label feature associated with their elements; these include natural language expressions that are meant to "provide a human-readable version of a resource's name"[3] and that act very often as domain specific terms.

It is an empirical issue whether linguistic and semantic analysis of the formal description and machine-readable representation of such labels would support the task of associating classes and properties of KR systems with (fragments of) textual documents. If an OBIE application detects information that corresponds to T-Box elements, this information can be marked as their related A-Box *instances*. Ontology Population (OP) then consists in storing all instances of taxonomy or ontology classes and properties we can extract from text in a knowledge base.

The work described in this paper is closely related to the "LexInfo' (Buitelaar et al.2009), (Declerck and Lendvai2010) and to the "lemon" (lexi-

[1] See also (Buitelaar et al.2008) for more details

[2] See (Baader2009) for more details.

[3] http://www.w3.org/TR/rdf-schema/#ch_label

con model for ontologies)[4] models that all work towards the goal of describing and representing lexical and linguistic properties of the textual content of taxonomy and ontology labels. On this basis, we started to analyze the textual content of labels encoded in XBRL taxonomies (see section 2 below) in order to see if this type of text can be used for supporting the task of finding corresponding information in related textual documents, like for example annual reports of companies. We discuss in detail some examples below after having briefly introduced the XBRL framework.

2 XBRL

XBRL, eXtensible Business Reporting Language[5], is an XML-based mark-up language for the exchange of business information, including financial reporting. XBRL specify the semantics of business data, its presentation, its calculation, and associated business rules, which are called formulas. XBRL also has its own special terminology and comes up in the form of a taxonomy, that is used for modeling various types of international standards[6] and national or regional legislations for financial reporting[7]. An XML document that contains concrete values for a number of XBRL concepts, like name of the company, period of the reporting and concrete values for financial items is called an instance document[8].

3 Examples of Terms in Labels and in Text

In section 3.1 four examples are given of textual content of labels in the IFRS taxonomy encoded in XBRL. Section 3.2 illustrates the typical content of a financial table of an annual report (in this case from the Deutsche Bank company, in German). In section 3.3 a short, partial segment of an explanatory note, in German, of a financial report (the company Bayer AG) is displayed.

It can be observed that neither the vocabulary of financial reports, nor the grammatical realizations of the concepts is harmonized with that used in labels. Our goal is to automatically assign the relevant concepts of the IRFS-XBRL taxonomy to (segments) of the two types of financial reports, and to transform (parts) of those documents onto an XBRL instance document with high precision.

3.1 Examples from the IFRS-XBRL Taxonomy

In each example below we have the name of the concepts (in italics within brackets) and both the corresponding English and German labels.[9]

1. Reconciliation of minimum finance lease payments payable by lessee / Überleitungsrechnung der vom Leasingnehmer im Rahmen von Finanzierungs-Leasingverhältnissen zu zahlenden Mindestleasingzahlungen (*ReconciliationOfMinimumFinanceLeasePaymentsPayableByLesseeAbstract*)

2. Reconciliation by end of reporting period / Überleitungsrechnung am Abschlussstichtag (*ReconciliationByEndOfReportingPeriodAbstract*)

3. End of period not later than one year / Bis zu einem Jahr bis zur Ende der Periode (*EndOfPeriodNotLaterThanOneYearAbstract*)

4. Minimum finance lease payments payable, at present value, end of period not later than one year / Im Rahmen von Finanzierungs-Leasingverhältnissen zu zahlende Mindestleasingzahlungen, zum Barwert, bis zu einem Jahr bis zum Ende der Periode (*MinimumFinanceLeasePaymentsPayableAtPresentValueEndOfPeriodNotLaterThanOneYear*)

3.2 Example from a Financial Table

```
Finanzleasingverpflichtungen
     275  25  46  60  144
```

This particular line is about the value of to be paid finance leases for the next periods: the total amount is 275 million euros and the periods are 1 year, 1-3 years, 3-5 years, more than 5 years.

[4] see: http://www.isocat.org/2010-TKE/presentations/Monnet-slides.pdf

[5] See http://www.xbrl.org/Home/

[6] Like the International Financial Reporting Standards (IFRSs), see http://www.ifrs.org/Home.htm

[7] For example the so-called General Accepted Accounting Principles (GAAP) of different countries, like Germany or the United States of America. The IFRS, the German and the US GAAPs, among others, can be browsed at http://www.abrasearch.com/ABRASearch.html

[8] Examples of these can be retrieved among others at the U.S. Securities and Exchange Commission (SEC, http://xbrl.sec.gov/) or at the Belgian National Bank (BNB, http://euro.fgov.be/.

[9] As an addtional information: The four concepts are in a sub-class relation in the taxonomy: $4 > 3 > 2 > 1$.

3.3 Example from an Explanatory Note

This (partially reproduced) note is describing the policy of the company with respect to finance leases.

"Ist der Bayer-Konzern Leasingnehmer in einem Finanzierungsleasing, wird in der Bilanz der niedrigere Wert aus beizulegendem Zeitwert und dem Barwert der Mindestleasingzahlungen zu Beginn des Leasingverhältnisses ... Die Mindestleasingzahlungen setzen sich im Wesentlichen aus Finanzierungskosten und dem Tilgungsanteil der Restschuld zusammen. ... Ist ein späterer Eigentumsübergang des Leasinggegenstands unsicher, ... Die zu zahlenden Leasingraten werden nach der Effektivzinsmethode aufgeteilt ... . Ist der Bayer-Konzern Leasinggeber in einem Finanzierungsleasing, werden in Höhe des Nettoinvestitionswerts Umsatzerlöse erfasst und eine Leasingforderung angesetzt. ...."

4 Our Approach to the Linguistic and Semantic Enrichment of Labels

We follow a multi-layered approach, starting with layout analysis, on the top of which linguistic and semantic analysis are proposed.

4.1 Segmenting and Tokenizing the Terms

In a first step, we segment the terms used in the labels (as listed in Section 3.1). For this one can make use of IFRS guidelines on the terminology used in the taxonomy, e.g. some punctuation signs explicitly mark term/sub-term segments (e.g. the commas segment term (4) in Section 3.1 into three subterms).

This approach is being consolidated by checking if the suggested sub-terms are themselves used as full terms in the labels of other concepts. In the given case we verify that this holds for only two subterms, but not for `zum Barwert` (*at present value*). From the linguistic point of view, we can tentatively associate the "consolidated" subterms with a status similar to an "arguments" of a functional term (to be established still).

4.2 Linguistic Analysis of the Terms

Subsequently, lemmatisation of the words used in the terms is performed in order to detect and link all possible forms of e.g. `Finanzierungs-Leasingverhätnissen` (*finance lease*) – its current inflection is dative plural, but the same term with other inflectional suffixes can be present in other labels of the taxonomy, or in external documents.

Next, we propose performing PoS tagging and complex morphological analysis, including derivation and compounding. This allows for example to detect in texts related terms such as `Finanzierungskosten` (occurring in the example in Section 3.3) and `Finanzleasingverpflichtungen` (occuring in the example of Section 3.2).

A chunking and a dependency analysis are also proposed, following the approach described in (Declerck and Lendvai2010), but refraining from showing the linguistic annotation due to limitations of space. Dependency analysis allows for detecting head nouns in terms. We can then compare labels sharing at least one identical head noun (its lemma) and thereby establish lexical semantic relations across concepts , taking into account the different linguistic contexts in all those labels.

Lemmas of head nouns are also considered as anchors for starting the search of relevant segments in textual documents. This strategy is motivated by the fact that in the taxonomy labels mainly nominal phrases are present.

4.3 Semantic Enrichment

Semantic annotation of subterms is recommended in case they represent temporal information (*end of reporting period*). Semantic enrichment can further be proposed on the basis of information that is either internal or external to the taxonomy.

An example for the internal case: as we noted in Section 3.1 the concept listed under (4) is a subclass of the concept listed under (2). We observe that none of the words used in the German label of the subclass occurs in the label of the superclass. But in both cases there is a subterm that can be annotated as a temporal expression (`Bis zu einem Jahr bis zur Ende der Periode` and `am Abschlussstichtag`). Between those expressions one can thus assume a semantic relation (the one containing in duration the other one, but we can also infer a lexical semantic is-a relation between *Minimum finance lease payments* and *Reconciliation*).

An additional semantic information we can in-

fer from internal information is about the semantic roles: the payments, which are a reconciliation, have a *lessee* and a *lessor*. This information is distributed over two classes, which are both at the (local) highest level in the taxonomy. This information helps to detect in text the corresponding concepts. But differently, depending if the document basis is a table or a free text. In the first case the semantic role *lessee* has to be infered as being the author of the document (the company providing for the annual report), since in tables nthe name of the company is normally not mentioned. In the second case both roles can be found, and here the use of Named Entity recognition tools is required.

With external enrichment we mean the use of resources like WordNet or FrameNet etc. for "importing" into the ontology labels additional lexical-semantic information.

We have to note here that with this issue the "classical" annotation of the terms with the means of XML, as proposed by (Declerck and Lendvai2010) comes to its limit. We plan therefore to test the lemon model[10] for encoding the linguistic and semantic enrichment of the labels of the taxonomy. It will be interesting to see if the resulting network of linguistic and semantic information, on the basis of the analysis of the "human-readable version" of the taxonomy is still comparable with the original concept-based taxonomy.

5 Conclusion and future work

We described in this short paper actual work on enriching taxonomy and ontology labels with linguistic and semantic information. With this approach we follow two goals: Improving the effectiveness and quality of ontology-based information extraction and possibly suggesting re-organizing the actual model of the domain of consideration.

In the case of XBRL taxonomies we see a large potential for getting not only a more compact but also a more complete model of the domain under consideration. While we are still using an XML annotation schema for this enrichment work, we plan to move to the RDF model proposed by lemon in order to support an ontological organization of the linguistic and semantic enrichment of the labels.

We are currently implementing a unification-based approach for comparing the linguistic and semantic features of the labels in KRs and of the result of the processing of the textual documents. This allows to make use of underspecification in the matching of information included in both sides, while requiring identity in the values of the "lemma" features.

We note finally that since the size of the taxonomy is limited and that many sub-terms are repeated in various concept labels, we can imagine a manually supervised annotation of the labels, this in order to ensure a high quality result of this task.

[10] As a reminder, see: http://www.isocat.org/2010-TKE/presentations/Monnet-slides.pdf

Acknowledgments

The ongoing research described in this paper is part of the RD project Monnet, which is co-funded by the European Union under Grant No. 248458 (see http://www.monnet-project.eu/). The conribution by Piroska Lendvai is co-funded by the European project CLARIN (www.clarin.eu).

References

Baader, F. (2009). Description logics. In *Reasoning Web: Semantic Technologies for Information Systems, 5th International Summer School 2009*, Volume 5689 of *Lecture Notes in Computer Science*, pp. 1–39. Springer–Verlag.

Buitelaar, P., P. Cimiano, A. Frank, M. Hartung, and S. Racioppa (2008). Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies* (11), 759–788.

Buitelaar, P., P. Cimiano, P. Haase, and M. Sintek (2009). Towards linguistically grounded ontologies. *The Semantic Web: Research and Applications*, 111–125.

Declerck, T. and P. Lendvai (2010). Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In *LREC 2010- The seventh international conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10), May 19-21, Valetta, Malta*. ELRA.

An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS

Claire Bonial* **Susan Windisch Brown*** **William Corvey*** **Volha Petukhova** **
Martha Palmer* **Harry Bunt****
*Department of Linguistics, University of Colorado at Boulder
**Department of Computer Science, Tilburg University
{Claire.Bonial,Susan.Brown,William.Corvey,Martha.Palmer}@colorado.edu
{V.Petukhova,Harry.Bunt}@uvt.nl

Abstract

This research compares several of the thematic roles of VerbNet (VN) to those of the Linguistic InfRastructure for Interoperable ResourCes and Systems (LIRICS). The purpose of this comparison is to develop a standard set of thematic roles that would be suited to a variety of natural language processing applications. Differences between the two resources are discussed, and in some cases, VN plans to adopt a corresponding LIRICS thematic role. In other cases, the motivations behind maintaining different thematic roles from those found in LIRICS are addressed.

1 Introduction

1.1 Motivation

The ideal set of thematic roles should be able to concisely label the arguments of any relation; however, what this set of roles should be has long been a subject of dispute in the linguistic community. In our current endeavor to create a possible standard set of thematic roles for the International Standards Organization (ISO), we have undertaken a systematic comparison of two semantic resources: LIRICS[1] and VerbNet (VN) (Schuler , 2002). We take a bottom-up approach in comparing roles across resources; the initial findings of this comparison follow, and demonstrate some of the difficulties in determining the ideal mapping between the thematic roles of LIRICS and VN.

1.2 Overview of LIRICS

The EU-funded project LIRICS was set up as a spin-off of ISO TC 37/SC4, with the aim of exploring the possibility of establishing sets of annotation concepts, defined in accordance with ISO standard 12620 as so-called data categories, for syntactic, morphosyntactic, and semantic annotation and lexical markup. In the part of the project concerned with semantic annotation, several approaches and existing annotation schemes for semantic role labeling were analyzed and compared with respect to (1) description model; (2) granularity; (3) definition of semantic roles; and (4) consistency and reliability of annotation (Petukhova and Bunt, 2008). Based on this study, it was concluded that semantic roles should be defined:

- as neither syntactic nor lexical structures but as semantic categories;
- by virtue of distinctive semantic properties;
- that are not restricted to only a few specific verb (noun, adjective) classes;
- as relational notions that link participants to an event, describing the way the participant is involved in an event (e.g. does he act intentionally; is he/it affected, changed, manipulated by other participants; does it come into existence through the event), rather than by internal properties).

A set of 29 semantic roles[2] was defined by listing for each a characteristic set of entailments. These entailments were converted into a set of orthogonal properties, e.g. [+/- intentionality], [+/- independent existence], etc. (see also (Dowty, 1991) and (Sowa, 2000)). For example, the *Theme* role is defined as a participant in an event or state who (i) is essential to the event taking place but does not have control over the way the event occurs; (ii) is not structurally changed by the event; in a state, is in a fixed position or condition throughout the state; (iii) is causally involved or

[1] Linguistic InfRastructure for Interoperable ResourCes and Systems http://LIRICS.loria.fr

[2] This set includes 11 roles which are central to any event, e.g. *Agent, Theme, Patient*; 10 adjunct roles, e.g. *Time, Location, Manner*; and 8 sub-roles for *Time* and *Location*, e.g. *Duration, Frequency, Path*. For definitions and examples of see http://let.uvt.nl/general/people/bunt/docs/LIRICS_semrole.htm

affected by other participants; (iv) in a state is essential to the state being in effect; but it is not as central to the state as a participant in the Pivot role.

Different levels of granularity are distinguished, where a low-level semantic role inherits all the properties of a high-level role and has an additional feature, which reflects additional or more specific entailments.

The LIRICS set of semantic roles was evaluated with respect to redundancy, completeness and usability for reliable, consistent annotation using a multilingual test suite including English, Dutch, Italian and Spanish (see (Petukhova and Bunt, 2008) and (Bunt et al., 2007)).

1.3 Overview of VN

The purpose of VN is to classify English verbs based on semantic and syntactic regularities; it has been used for numerous NLP tasks, most notably, semantic role labeling ((Schuler , 2002) and (Shi and Mihalcea, 2005)). In each verb class, the thematic roles are used to link syntactic alternations to semantic predicates, which can serve as a foundation for further inferencing. For this reason, VN relies to an extent on syntactic features. Because VN is organized into verb classes, it is desirable to have an explicit hierarchy of roles such that users can understand the specificity of a role for a given class, as well as the superordinate category of that role, which would apply to classes of verbs that take diverse arguments. For example, the VN role *Topic* is a type of *Theme* that is restricted to arguments that express the transfer of information. The specificity of this role helps distinguish certain classes of verbs from others, and its compatibility with a particular verb helps determine whether that verb belongs in a certain class. However, thematic roles alone do not determine class membership; rather, a verb's thematic roles are considered along with the verb's semantics and syntactic patterning in assigning the verb class. The use of roles that are specific to certain classes of verbs is informative for VN users: roles that are unique to a particular class of verbs are maximally specific in their characterization yet amenable to hierarchical arrangement, which allows users to assign roles at various levels of granularity.

2 Thematic Roles in Comparison

LIRICS and VN thematic roles largely overlap; however, the divergent goals and structures of the resources occasionally yield different roles. Attempting to find the ideal mapping between the roles of these two resources will be a first step in establishing an optimal set of standard thematic roles: roles that can generalize across the greatest number of syntactic and pragmatic contexts, while bringing the most appropriate level of specificity when naming an event participant. The following sections detail some of the challenges discovered in our initial comparisons of VN and LIRICS semantic roles.

2.1 VN *Actor1, Actor2* vs. LIRICS *Agent, Partner*

Verbs such as *chat, cooperate,* and *speak* correspond to events that usually involve two volitional participants, as in: 'Susan chatted/cooperated with Rachel.' Currently, VN uses the labels *Actor 1* and *Actor 2* to refer to each of these participants. In typical usage, *Actor 1* is the subject of the verb and *Actor 2* occurs in the oblique (e.g. 'with Rachel'). In theory, these labels capture the notion of two volitional actors involved in a single event, where one seems to be a true agent with pragmatic focus (*Actor 1*), while the other participant (*Actor 2*) fulfills the same agentive qualities (animate, volitional) without pragmatic focus.

While LIRICS does not have an exact mapping to *Actor 1* and *Actor 2*, it does have the complementary roles of *Agent* and *Partner*. In the LIRICS framework, an *Agent* is defined as a 'participant in an event who initiates and carries out the event intentionally or consciously, and who exists independently of the event,' while a *Partner* is defined as a 'participant in an event who is intentionally or consciously involved in carrying out the event, but who is not the principal agent of the event, and who exists independently of the event.' Upon examining this distinction between *Agent* and *Partner*, we decided that we preferred the LIRICS terms for the following reasons: 1) the labels *Agent* and *Partner* more clearly indicate that there are differing levels of agency between the two roles 2) using the term *Actor 1* fails to illustrate that the argument is essentially an agent.

2.2 VN *Theme 1, Theme 2* vs. LIRICS *Theme, Pivot*

Unfortunately, an adoption of *Agent* and *Partner* produces a potentially confusing incongruency among VN roles: parallel to *Actor 1* and *Actor 2*, VN has the roles *Theme 1, Theme 2,*

Patient 1 and *Patient 2*. *Theme 1* and *Theme 2*, for example, are used for verbs such as *border, coincide*, and *have*, which denote events that may involve two themes: 'Italy-*Theme 1* borders France-*Theme 2*.' The relationship between the two themes is analogous to the relationship between *Agent* and *Partner*: there is a pragmatically focused theme (Theme 1) and a secondary theme (Theme 2). In order to accommodate this parallelism, we decided to maintain the concept behind LIRICS *Agent* and *Partner*, but adjust the labels to *Agent* and *Co-Agent, Theme* and *Co-Theme* and *Patient* and *Co-Patient*. Perhaps more importantly, there is the rare possibility that a sentence could involve both a *Partner* to an *Agent* and a *Partner* to a *Theme* or *Patient* leading to two ambiguous *Partner* arguments; the 'Co-' terminology allows them to be easily distinguished.

In certain cases, we found that the role *Theme 1* could be better expressed using another LIRICS role: *Pivot*, a 'participant in a state that is characterized as being in a certain position or condition throughout the state, and that has a major or central role or effect in that state.' For verbs in the Own and Require classes, using *Theme 1* to refer to the possessor or requirer seemed to obscure an important distinction between this type of participant and other *Theme 1* arguments, wherein the *Theme 1* is primarily being located (e.g. 'Italy' in 'Italy borders France'). For verbs in the Own and Require classes, *Theme 1* is not located; instead it is involved in a state of ownership or need. Thus, for the Own and Require classes, we did not adopt *Theme* and *Co-Theme* to replace *Theme 1* and *Theme 2*; rather, we chose to adopt the label *Pivot* for participants in a state of ownership or need, and to use *Theme* to refer only to the owned or needed participants. We expect to utilize *Pivot* in similar circumstances throughout the resource.

2.3 VN *Topic* vs. LIRICS *Theme*

In VN, *Theme* is used with a wide variety of verbs to label a participant that is being literally or metaphorically located, positioned, or moved; this participant may be concrete or abstract. *Topic*, on the other hand, is restricted to participants involved in the transfer of information: arguments of verbs such as *advise, promise*, and *tell*. For example: 'John-*Agent* informed me-*Recipient* of the situation-*Topic*.' *Topic* inherits all of the features of *Theme*, but is constrained by additional features such as +*information content* and +*abstract*. LIRICS does not use *Topic*, instead *Theme* would be used for these arguments. This discrepancy illustrates the differing aims of the two resources: VN uses finer-grained roles where this can help to distinguish classes, a practice LIRICS specifically avoids.

2.4 VN *Stimulus, Experiencer* vs. LIRICS *Cause, Pivot*

Verbs such as *see, amuse* and *empathize* involve one participant that is perceiving another in a cognitive or sensory manner, but the event does not necessarily involve contact or volition on the part of either party. The participant that triggers the event does not do so purposefully; the fact of its existence, perceived by another participant, yields the physical or mental reaction in that participant. In the LIRICS framework, the trigger would be a *Cause*, defined as a 'participant in an event (that may be animate or inanimate) that initiates the event, but that does not act with any intentionality or consciousness; it exists independently of the event,' while the participant reacting would be *Pivot*. In VN, the trigger is inconsistently labeled either *Cause* or *Stimulus*, while the participant reacting is the *Experiencer*. After an examination of all of the verb classes using these roles in combination (27 classes), we defined *Stimulus* as a participant that unintentionally arouses a mental or emotional response in a sentient being. In turn, we found that the *Experiencer* was consistently a participant undergoing a particular mental or emotional state precipitated by the mere perception of another participant. Thus, we found that *Stimulus* and *Experiencer* emerged as a natural pairing in verb classes involving a cognitive or emotional event. *Stimulus* is thus a more constrained type of *Cause*, where the causation is mediated by cognitive experience: 'The storm-*Stimulus* frightened the children-*Experiencer*,' vs. 'The storm-*Cause* destroyed the ship-*Patient*.' Therefore, although we find the LIRICS roles of *Pivot* and *Cause* to be very useful and have their place within VN, we also believe that the greater specificity of the *Experiencer* and *Stimulus* roles, which helps to distinguish verb classes, should be maintained.

2.5 Discussion: Remaining open issues

As demonstrated in the comparisons presented, decisions concerning one thematic role often impact other thematic roles and thematic role pat-

terns across the resource. For this reason, it is important to keep multiple thematic roles in mind when analyzing the impact of proposed changes. Further, this paper has explored possible subset relations among thematic roles, raising questions about the nature and depth of hierarchical relationships among roles. For instance, the LIRICS role *Goal* corresponds to VN *Recipient*, and the LIRICS *Final Location* corresponds to VN *Destination* (Petukhova and Bunt 2008). However, in many semantic frameworks, *Goal* represents an end location of an action that would subsume both *Recipient* and *Destination*. In examining possible benefits of incorporating *Goal* into a hierarchy of VN roles, we confront questions about congruency of scope (in this case, *Goal* versus *Final Location*) between semantic roles at a given level within a hierarchy.

Analysis of several VN roles is still underway. As we have begun to show above, the roles *Source, Location, Destination,* and *Recipient* are closely related to each other, and an initial look into their use in VN suggested that the definitions should be clarified and additional roles considered. For example, the *Destination* role was initially used for goals that were physical locations but had been extended as new classes were added to include non-locative goals. In addition, its use seemed to overlap in some cases with the role *Location*. Our comparison with the LIRICS roles *Location, Initial Location, Final Location, Source, Goal,* and *Recipient* is contributing to our construction of explicit definitions for these VN roles and our consideration of new roles. The LIRICS features concerning the temporality and physical locality of these roles is helping direct our analysis. Additionally, the status of the role *Proposition* is in question, as its distinction from *Topic* may be purely syntactically motivated (i.e. *Propositions* only occur as clausal arguments). A summary of the role comparisons completed in the present study appears in Table 1.

Current VN role	LIRICS role	Proposed VN role
Actor 1	Agent	Agent
Actor 2	Partner	Co-agent
Patient 1	Patient	Patient
Patient 2	Partner	Co-patient
Theme 1	Pivot	Theme
Theme 2	Theme	Co-theme
Theme 1	Pivot	Pivot
Theme 2	Theme	Theme

Table 1: Summary of VerbNet role changes based on comparison to LIRICS (note that the exact *Theme 1* and *Theme 2* changes will depend upon the verb class under consideration).

2.6 Conclusion and Future Work

In this comparison process we are re-evaluating VN roles, allowing us an opportunity to create a clear definition for each role and to make changes ensuring that each role is used consistently throughout VerbNet; these definitions and changes are forthcoming. Our ongoing comparison of VN and LIRICS has demonstrated that resources differing in aim and structure can still overlap a great deal in their definitions of core thematic roles (e.g., *Agent, Patient, Instrument,* and to a large extent, *Theme*). Their differing goals can also result in some variation in their final sets of thematic roles and the boundaries of those roles. Although this highlights the difficulties that are involved in creating an ISO standard set of thematic roles, the process of comparison has also made explicit the motivations behind certain differences. In some cases, the comparison led to a revision of VN roles (e.g., adopting *Pivot* in place of *Theme* in certain situations, and changing *Actor 1* and Actor 2 to *Agent* and *Co-agent*), whereas in others the comparison helped develop more rigorous role definitions (e.g., for *Experiencer* and *Stimulus*). With a clearer understanding of the resources' motivations for the roles, we are better able to devise a set of thematic roles that are suited to the widest range of purposes.

Future work will perform similar comparisons over additional resources, notably FrameNet (Fillmore and Baker, 2010), as LIRICS already defines links to roles in this resource. The final goal of all comparisons will be the development of a set of thematic roles that is suited not just to the idiosyncratic purposes of one resource, but rather to a wide variety of natural language processing purposes. As we have shown in these initial comparisons, one of the most difficult issues in developing a standard resource compatible with different purposes is the issue of granularity, or the extent to which thematic roles are illustrative of different classes of verbs as opposed to generalizable across all verbs. To overcome this difficulty, we are adjusting VN such that the resulting thematic role

set will be hierarchical. A hierarchical structure with clear mappings between higher and lower-order classes will allow users to select the level of granularity that is best suited to the application they are developing, including only those thematic roles that generalize across all classes of verbs (e.g. *Theme*), or including more specific roles that are characterized by additional features, and therefore only appear in certain classes of verbs (e.g. *Topic*).

Research in automatic semantic role labeling, for example, has demonstrated the importance of the level of granularity of semantic roles. Yi, Loper and Palmer (2007) and Loper et al. (2007) both demonstrate that because VN labels are more generalizable across verbs than PropBank (Palmer et al., 2005) labels, they are easier for semantic role labeling systems to learn. However, Merlo and Van Der Plas (2009) found that the differing levels of granularity of PropBank and VN were both useful, and therefore suggest complementary use of both resources. Our hope is that the final set of thematic roles we decide upon, informed by our comparisons to other resources such as LIRICS and FrameNet, will encompass the benefits of hierarchical granularity, thereby meeting the unique needs of varying natural language processing applications.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grant NSF-0415923, Word Sense Disambiguation, and a DARPA supplement to a grant from the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, a subcontract from BBN, Inc.

References

Bunt, H., Petukhova, V., and Schiffrin, A. 2007. *LIRICS Deliverable D4.4. Multilingual test suites for semantically annotated data*. http://lirics.loria.fr.

Dowty, D. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67:547-619.

Fillmore, C.J. and Baker, C.F. 2010. A Frame Approach to Semantic Analysis. In Heine, B. and Narrog, H. (eds.) *Oxford Handbook of Linguistic Analysis*: Oxford University Press.

Kipper, K. 2002. VerbNet: A Class-Based Verb Lexicon. http://verbs.colorado.edu/mpalmer/projects/verbnet.html

Loper, E., Yi, S., and Palmer, M. 2007. Combining Lexical Resources: Mapping Between PropBank and VerbNet. Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7), Tilburg.

Merlo, P., and Van Der Plas, L. 2009. Abstraction and Generalization in Semantic Role Labels: PropBank, VerbNet or both? Proceedings of the 47th Annual Meeting of the ACL and 4th IJCNLP of the AFNLP, Suntec, pp. 288-296.

Palmer, M., Gildea, D., and Kingsbury, P. 2005. The Proposition Bank: An annotated Corpus of Semantic Roles. Computational Linguistics, 31:1, pp. 71-105.

Petukhova, V., Schiffrin, A., and Bunt, H. 2007. Defining Semantic Roles. Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7), Tilburg, pp. 362-365.

Petukhova, V., and Bunt, H. 2008. *LIRICS semantic role annotation: Design and evaluation of a set of data categories*. Proceedings of the sixth international conference on language resources and evaluation (LREC 2008), Paris: ELRA.

Schuler, K.K. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon* University of Pennsylvania.

Shi, L. and Mihalcea, R. 2005. *Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing*. *Computational Linguistics and Intelligent Text Processing*, 100-111.

Sowa, J.F. 2000. *Knowledge representation: logical, philosophical and computational foundations*. Pacific Grove: Brooks/Cole.

Swier, R.S. and Stevenson, S. 2004. Unsupervised semantic role labelling *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 95-102.

Yi, S., Loper, E., and Palmer, M. 2007. Can semantic roles generalize across genres? Proceedings of the HLT/NAACL-2007, Rochester, pp. 548-555.

Classification and Deterministic PropBank Annotation of Predicative Adjectives in Arabic

Abdelati Hawwari[2], **Jena D. Hwang**[1,2], **Aous Mansouri**[1] **and Martha Palmer**[1,2]
[1]Department of Linguistics
[2]Center for Computational Language and EducAtion Research
University of Colorado at Boulder
Boulder, CO 80309
{abdelati.hawwari,hwangd,aous.mansouri,martha.palmer}@colorado.edu

Abstract

In Arabic, adjectives can occur as the predicating elements in sentences without verbs. Thus in the Arabic PropBank, we annotate these predicate adjectives similarly to verbs to provide meaningful representations for verbless sentences. In this paper, we present our analysis of the predicate adjectives and a framework for deterministically annotating their argument structure that approaches the accuracy of human annotators, which will be helpful in speeding up and expediting our annotation process.

1 Introduction

Predicate argument structures have long been considered useful representations for sentences. In Arabic, predicates can occur in sentences that contain a verb phrase and a verb or in a sentence that entirely lacks a verb. These latter type of sentences, which we will refer to as either "zero-copula" or "equative" sentences, tend to occur in present tense situations where a copula is not needed as seen in the following example, where the role of the predicate is satisfied by an adjective.

(1) Hind **mujtahidatun**
Hind **hard-working.nom**
'Hind (is) a hard-worker'

Thus in PropBank, we have decided to treat these predicate adjectives as we do verbs and annotate them with their semantic frames. This analysis will help us determine the arguments of all types of predicates, including adjectives; describe semantic relationships within certain compound noun phrases; and simplify the issue of translating nominal sentences in Arabic to languages that don't allow for zero-copulas.

This paper also proposes a framework for analyzing and automatically annotating the argument structure of predicate adjectives in Arabic. Our process approaches the accuracy of human annotators. This can speed up annotation by using the deterministically annotated data as one of the two annotations in a double-blind annotation process.

1.1 Arabic Treebank and PropBank

Arabic Treebank 3 (ATB3) is a 400K word corpus of Lebanese newspaper corpora *an-Nahar*. ATB3 annotates the syntactic structure of sentences the newswire corpora, following the guidelines, when possible, set by the Penn English Treebank (Maamouri et. al., 2009).

Referencing the syntactic parse provided by ATB3, Arabic PropBank annotates the semantic argument structure of predicates. The data is intended for use as training for automatic semantic labeling. Arguments are tagged based on their semantic relationship with the predicate - regardless of their syntactic relationship. For example, *the vase* receives the patient (Arg1) even it appears in different syntactic positions as in the following sentences: *The vase was broken* and *The cat broke the vase*. PropBank includes both the numbered arguments (e.g. Arg0 for agents/experiencers; Arg1 for patient/theme) and the adjuncts or modifiers (e.g. ArgM-TMP for temporals). In addition to English and Arabic, PropBanks also exist for Chinese (Xue and Palmer, 2009), Hindi (Palmer et. al., 2006) and Korean (Palmer et. al., 2009).

2 Background on Predicate Adjectives

2.1 Linguistic Analysis

In Arabic, "adjective" is not a part-of-speech (POS) category of its own. Traditionally, grammarians have identified three POS categories, which are based on the lexical items' morphology: verbs, nouns, and particles. The last category is a closed set of functors (e.g. conjunctions, pronouns, etc). Verbs are based on triconsonantal roots and follow strict patterns consisting of specially ordered consonants and vowels in their conjugations and inflections. The rest fall into the noun category.

The noun category is a rather large category that includes nouns, proper nouns, and participles. What we would consider adjectives in English, including descriptors such as *big* or *small*, are amongst the Arabic nouns, or participles, to be exact. Thus, Arabic does not have adjectives *per se*, rather the status of "adjective-ness" is a function of how the participles behave in their syntactic environment.

Traditional Arabic grammarians have also recognized that nominal participles display syntactic behavior that imitates that of a verb. Adopting the subject/predicate paradigm from the standard Greek ideas of grammar (El-rajehy, 1979), they were aware of the prominence of the verb in a clause. As for equational sentences which lack an overt verb, they studied what is most similar to a verb. Subsequently in an attempt to draw a parallel between the sentences with a verb and those without a verb, they divided the clause into two parts: subject and predicate. This division of equational sentences was supported by a number of factors: i) most nominal predicates are actually derived from verbs, ii) for cases where the predicates are not derived from verbs, one can substitute a derived nominal to complete the structure (2), iii) in cases where the predicate is a PP, one can infer a predicate (3). Predicate adjectives are in bold in the examples.

(2) ar-rajulu wa l-mar'atu sawā' **/ mutasāwūna**
man.nom conj woman.nom same **/ equal**
'Man and woman (are) the same / equal'

(3) Hind fī l-manzili >**mawjūdatun** fī l-manzili
Hind in home >**exist** in home
'Hind (is) in the house >exists in the house'

	N	Participles			V
Syntactic Behavior		verb-like > <noun-like			
Traditional Analysis	Nouns				Verbs
ATB3		ADJ	ADJP-PRD	ADJ. VN	

Table 1: Defining Predicate Adjectives.

Effectively, what was recognized was that the nominal participles can take the role of a predicate in a equative or zero-copula sentences, and are called predicate adjectives here. In other words, predicate adjectives are traditionally categorized nouns and behave syntactically like verbs.

Despite the recognition of the syntactic behavior of these predicate adjectives, as discussed here, the vast majority of the linguistic literature has traditionally swept the predicate adjectives into the noun category and focused mostly on the more noun-like participles such as descriptive adjectives (e.g. *big* or *small*). Even contemporary linguists, such as Kremers (2003), continue to analyze adjectival predicates as a subtype of nouns based on morphological analyses and this explains a lack of attention being paid to the adjectival predicates.

2.2 Predicate Adjectives

If we were to rearrange the POS based on syntactic functionality, we would have the nouns on one side of the spectrum and the verbs on the other side (Table 1), with a grey zone in the middle where these participles would lie. Some participles would be placed closer to the noun side and more verb-like participles would be placed near the verbal end of the spectrum The *hard-working* in the example (1) would be closer to the verb side of the spectrum as it is the predicating participle in the absence of a verb, while the same participle in example (4) is closer to the noun as it is a nominal modifier for *student*.

(4) Hind ṭālibatun **mujtahidatun**
Hind student.nom **hard-working.nom**
'Hind is a hard-working student'

These verb-like participles do not only act as predicates, but display another verbal quality which

is gives the direct object accusative case as seen in example (5).

(5) Maḥfūẓ **ḥāmilun** jā'izata nōbil
Mahfouz **reciever.nom** award.acc Noble
'Mahfouz (is) a receiver (of the) Noble Awards'

Here, the predicate adjective 'receiver' is not only acting as the predicate of the clause, it is also syntactically functioning as a verb by giving *award* accusative case.

2.3 Arabic TreeBank

Morphologically, ATB3 has classified the POS of the adjectives based on traditional classification. However, within the syntactic parse, ATB3 makes a three-way distinction amongst the participles based on their syntactic behavior. ATB3 treats the participles as nouns if they act as a modifier in a noun phrase, much like English adjectives. If the participles are found in equational sentences with zero-copula where the participle is a predicate (i.e. case of predicate adjective), ATB3 places the predicate within an adjectival phrase marked as predicate, namely ADJP-PRD. Finally, if an Active or Passive participle (see section 3 for description) is found in a zero-copula sentence with the ability to assign accusative case marker to its object, it takes a verbal reading and the participle receives the tag of ADJ.VN.

3 Classification of Predicate Adjectives

We analyzed all ADJP-PRDs as marked by ATB3 and each adjective was manually classified into one of 7 classes. Our grouping follows ATB3's adjectival and participial classes mentioned in their guidelines (Maamouri et. al., 2009) - keeping in mind that these distinctions were not marked on the trees themselves. Both the ATB3's and our classifications closely adhere to traditional categorizations of these participles (Hassan, 1974), which are based on morpho-syntactic features unique to each class.

Relational adjectives (RA) are derived from nouns. The resultant word references the noun from which it was derived (e.g. 'Egypt' *miṣr* >'Egyptian' *miṣriyy*). We found 253 tokens and 122 types of RA in our data.

Comparative adjectives (CA) are derived from stative verbs and are used in comparing attributes between different items (e.g. 'be good' *ḥasuna* >'better/best' *'aḥsan*). We found 14 tokens and 7 types of CA in our data.

Exaggeration participles (EPA) are used to ascribe to an entity an exaggerated description (e.g. 'to lie' *kaḏaba* >'liar' *kaḏḏāb*). We found 22 tokens and 16 types of EPA in our data.

Inherent participles (IPA) are derived from stative verbs and indicate a stable inherent feature (e.g. 'to become red' *iḥmarra* >'red' *'aḥmar*). We found 328 tokens and 101 types of IPA in our data.

Active participles (APA) describe the agent and are derived from active (non-stative) verbs (e.g. to teach' *darrasa* >'teacher' *mudarris*). We found 716 tokens and 297 types in our data.

Passive participles (PPA) describe the patient and are derived from active (non-stative) verbs (e.g. 'to write' *kataba* >'something written' *maktūb*). We found 408 tokens and 157 types in our data.

Miscellaneous Adjectives (MA) are mostly elements that the ATB3 has marked as an adjective due to their syntactic behavior. They are a closed list of words that have been sub-classified into 3 groups based purely on their argument structure. We found 3 tokens and 1 type of MA in our data.

4 Deterministic Annotation

From the seven classes we have identified, we evaluated the syntactic composition of each of the classes as parsed by the gold annotation of ATB3 to determine if a set of rules can be derived for deterministic annotation.

We determined that classes RA, CA, and IPA would lend themselves to deterministic identification of the annotation of the argument structure. We also determined that MA would be a good class for deterministic annotation; however since we had a very small number of these, we decided to skip this class until we have further data.

The rest of the classes, we found, do not lend themselves to deterministic identification due to their close identification with the verbs they are derived from. This causes ambiguities because of multiple senses and syntactic variations of the lemma. Due to the limitations of this paper we can only give

Class	if Sisters of Predicate	if Sisters of ADJP-PRD
RA	If immediate next sister is NOUN, tag as ARGM-EXT If immediate next sister is PP, tag as ARG2 Else: ARGM-XXX if function tag XXX in ATB3 exists, else ARGM-ADV	If NP-SBJ is a sister to ADJP-PRD, then it is ARG1 Else first NP is ARG1. ARGM-XXX if function. tag XXX in ATB3 exists, else ARGM-ADV
CA	Does definite exist? If yes: NP next to predicate is ARG3 If not: NP next to predicate is ARG2 and second NP is ARG3 Else: ARGM-XXX if function tag XXX in ATB3 exists, else ARGM-ADV	
IPA	If exists NP or PP and immediately next to predicate, tag as ARG2 If exists NP or PP and no ARG2 specified yet, tag as Arg2 If NP-ADV or NOUN, tag as ARGM-EXT Else: ARGM-XXX if function tag XXX in ATB3 exists, else ARGM-ADV	

Table 2: Heuristics used for deterministic annotation of RA, CA, and IPA.

one ambiguous example, example (6) below.

(6) al-quwwāt al-**muḥtalla** wa n-niẓām
l-'irāqiyyu [...] hiya l-mas'ūla 'an [...] 'asrā
al-ḥarbi fī l-'arāḍi l-muḥtalla

det-forces det-**occupy** conj. det-regime
Iraqi [...] she det-responsible [...] for
captives det-war in det-lands det-occupy

'The **occupying** *forces and the Iraqi regime, they are responsible for prisoners of war in the occupied lands.'*

Here due to phonological processes that merge the underlying forms into a single actualized word, the APA (bolded) and PPA (underlined) for the predicate *muḥtalla* share the same lexical shape.

4.1 Heuristics and Results

For the classes RA, CA, and IPA, we used simple heuristics that generalize over the entire class of adjectives (see Table 2). In PropBank we limit the scope of our annotation to the sentence in which the predicate appears. We kept the same scope of annotation for predicate adjectives: we annotated the nodes that are sisters of the predicate adjectives and sisters to the ADJP-PRD. If there were more than one predicate adjective connected through conjunction, each predicate adjective was annotated separately. In addition to this, the following nodes were skipped PRN: "parentheticals", NAC: "not a constituent", and PRT: "particles" often referring to functional words (e.g. conditionals, conjunctions, etc). We experimented with 481 sentences from ATB. This translates into 595 instances or predicate tokens and 230 predicate types.

For evaluation, we took the inter-annotator agreement (ITA) between the deterministically tagged data and the same data annotated by a linguist specializing in Arabic Lexicography. When we evaluate our data based on the evaluation method for CoNLL's semantic role labeling task (Carreras and Màrquez, 2005) – exact match on labels, the agreement for the annotation is 91.6% for IPA, 89.9% for RA, and 84.6% for CA. If the function tags on ARGM (e.g. TMP in ARGM-TMP) are ignored, then the agreement rates is 92.5% for IPA, 93.2% for RA and 84.6% for CA.

When manually examining the data we noticed that all errors for the predicate adjective type CA were the case of missing ARGMs, which accounts for why ignoring the function tag did not help the ITA at all. For IPA and RA, there were no obvious patterns in the mistakes our deterministic system was making. However, we should note that in all cases, these results are at least as good as the PropBank agreement rate between two human annotators (Kingsbury and Palmer, 2002)[1].

[1]The interannotator agreement results in (Kingsbury and Palmer, 2002) are based on English PropBank data. Though we are working on Arabic with different syntactic structures, we still are analyzing predicating structures. That is, just as verbs are predicates of sentences in English, so are adjectives in Arabic's verbless sentences.

5 Conclusion

In this paper, we have presented our work on expanding predicate argument annotation in APB by including the predicative adjectives and a deterministic annotation of a subset of these predicate adjectives. Our aim was to develop the a method for semi-automating annotation for quicker and more efficient annotation. Our results show that deterministically annotated data approaches the accuracy of human annotators and would be helpful in speeding up and expediting our annotation process. Along these lines, further work on nominal predicates in APB will include framing and annotating other participles, in addition to the ADJP-PRDs, such as ADJ.VNs.

Acknowledgements

We are grateful to Mona Diab and Wajdi Zaghouani for helpful discussions and comments. We also acknowledge the support of the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No HR0011-06-C-0022, subcontract from BBN, Inc.

References

Carreras, Xavier and Lluís Màrquez. 2005 Introduction to the CoNLL-2005 shared task: semantic role labeling. *CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning.* Ann Arbor, Michigan. June 30, 2005.

El-rajehy, Abduh. 1979. *Arabic Syntax and Modern Linguistics: A Study in Methodology.* 62

Kingsbury, Paul and Martha Palmer. 2002. From Treebank to PropBank. *Third International Conference on Language Resources and Evaluation.* LREC-02., May 28- June 3, 2002, Las Palmas, Canary Islands, Spain: 2002.

Kremers, Joost. 2003. Adjectival Agreement in the Arabic Noun Phrase. In *Proceedings of ConSOLE XI.* Padova, Italy, 2003.

Hassan, Abbas 1974. *al-Nahw al-Wafi.* Cairo: Dar al-Maarif.

Maamouri, Mohamed, Ann Bies, Sondos Krouna, Fatma Gaddeche, Basma Bouziri. 2009 *Penn Arabic Treebank Guidelines Version.* Version 4.92. http://www.ldc.upenn.edu/ wajdiz/treebank/annotation/syntax.html

Palmer, Martha, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. LDC Catalog LDC2006T03.

Palmer, Martha, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure. In *The 7th International Conference on Natural Language Processing* (ICON-2009). Hyderabad, India, Dec 14-17, 2009.

Xue, Nianwen and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering.* 15, 1 (Jan. 2009), 143-172.

Towards Interoperability for the Penn Discourse Treebank

Nancy Ide
Department of Computer Science
Vassar College
ide@cs.vassar.edu

Rashmi Prasad
Institute for Research in Cognitive Science
University of Pennsylvania
rjprasad@seas.upenn.edu

Aravind Joshi
Institute for Research in Cognitive Science
University of Pennsylvania
joshi@seas.upenn.edu

Abstract

The recent proliferation of diverse types of linguistically annotated schemes coded in different representation formats has led to efforts to make annotations interoperable, so that they can be effectively used towards empirical NL research. We have rendered the Penn Discourse Treebank (PDTB) annotation scheme in an abstract syntax following a formal generalized annotation scheme methodology, that allows meaning-preserving mappings to any other scheme. As an example, we show the mapping of the PDTB abstract syntax to a representation in the GrAF format.

1 Introduction

The last decade has seen a proliferation of linguistically annotated corpora coding many phenomena in support of empirical natural language research – both computational and theoretical. Because the annotated phenomena and annotation representations vary widely across different schemes, there is a need for making them compatible with each other, to ensure effective merging, comparison and manipulation with common software.

Ensuring compatibility is even more necessary when different types of annotations are done on the same source text, for example the *Wall Street Journal* (*WSJ*) corpus. Multi-level annotations on the *WSJ* include part of speech tagging, syntactic constituency, coreference, semantic role labeling, events, and discourse relations. In many cases, empirical natural language research using the *WSJ* would benefit immensely from using information from multiple layers of annotation, but in order to allow for this, it is imperative to ensure the efficient interoperability of the annotations.

In Bunt (2010), an annotation scheme design methodology is proposed, providing a formal specification of a representation format as a rendering of conceptual structures defined by an *abstract syntax*. It ensures that every representation format is convertible through a meaning-preserving mapping to any other representation format. In Ide and Bunt (2010), Bunt's design methodology is further generalized, and a mapping strategy is defined to convert from an abstract syntax to a representation in GrAF format (Ide and Suderman, 2007). To illustrate the process, Ide and Bunt generate an abstract syntax and apply the mapping strategy to annotation schemes for ISO-TimeML (ISO, 2009), PropBank (Palmer et al., 2005), and FrameNet (Baker et al., 1998).

In this paper, we present an application of this methodology to the annotation scheme of the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), which contains annotations of discourse relations on the *WSJ*. Our goal is to allow for effective combination of PDTB with GrAF renderings of PropBank and other annotations that have been done on all or parts of the *WSJ*, including Penn Treebank (PTB) syntactic annotations. In addition, we hope that this work will feed the development of a standard for annotating discourse relations in ISO project 24617-5 (Semantic Annotation Framework, Part 5: Discourse Relations).

2 The PDTB: Brief Overview

The PDTB (Prasad et al., 2008) provides annotations of discourse relations, along with their arguments, senses and attributions, on the entire PTB-II portion of the *WSJ* corpus, consisting of approximately 1 million words.[1] Some PDTB annotations are illustrated in Exs.(1-5). Discourse relations, such as causal, contrastive, and temporal relations, are triggered by explicit words or phrases (e.g., the underlined expressions in Exs. (1) and

[1] http://www.seas.upenn.edu/~pdtb. The corpus is distributed via LDC (http://www.ldc.upenn.edu).

(3), or by adjacency. Explicit realizations can occur via grammatically defined *connectives* (Ex. 1), or with other grammatically non-conjunctive expressions called *Alternative lexicalizations* (AltLex) (Ex. 3). The two arguments of a discourse relation are abstract objects (AO) in discourse, such as events, states, and propositions, and are labelled Arg1 (shown in italics) and Arg2 (shown in bold). Between two adjacent sentences not related by an explicit connective or AltLex, an implicit discourse relation can be inferred, when the annotator has to *insert* a connective to express the inferred relation (e.g., the implicit connective *because* inserted in Ex. 2). It is also possible for adjacent sentences to *not* be related by a discourse relation, in particular when the sentences are linked by an entity-based coherence relation (EntRel, shown in Ex. 4), or are not related at all via adjacency (NoRel, shown in Ex. 5). For each discourse relation, a sense (shown in parentheses at the end of examples), drawn from a hierarchical sense classification scheme, is provided for the relation. The attribution (to an agent of the AO assertion, belief, fact, or eventuality) of each discourse relation and each of its two arguments is also annotated, along with the attribution text when it is explicit (e.g., the attribution over Arg1 in Ex. 3).

1. *Big buyers like P&G say there are other spots on the globe, and in India, where the seed could be grown. . . .* But **no one as made a serious effort to transplant the crop**. (Comparison:Concession:Contra-expectation)

2. *Some have raised their cash positions to record levels.* Implicit=because **High cash positions help buffer a fund when the market falls**. (Contingency:Cause:Reason)

3. *But a strong level of investor withdrawal is much more unlikely this time around,* fund managers said . **A major reason is that investors already have sharply scaled back their purchases of stock funds since Black Monday**. (Contingency:Cause:Reason)

4. *Pierre Vinken, . . ., will join the board as a nonexecutive director Nov. 29.* EntRel **Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group**.

5. *Jacobs is an international engineering and construction concern.* NoRel **Total capital investment at the site could be as much as $400 million,** . . ..

PDTB annotations are stand-off, in that files containing the annotations are physically separate from the source text files. The PDTB annotation scheme and representation are fully described in the manual (PDTB-Group, 2008).

3 A Formalization of PDTB annotations

The current scheme for annotating a *discourse relation entity* in the PDTB includes a list of values, vertically represented in the annotation files. Values also represent *text spans*, as references to the character offsets in the source text file, and the *PTB alignments* of the text spans, as gorn address (Gorn, 1965) references to nodes in their corresponding PTB constituency trees. The full set of features and descriptions of their value assignments is given in Table 6.

The vertical representation of the PDTB annotations can also be converted to a simpler horizontal format, with each line corresponding to one discourse relation.[2] For this work, we have used the horizontal format in which the values for each field are separated by vertical bars; for example:

```
Explicit|258..262|once|1,0,1,2,0|Ot|
Comm|Null|Null|361..377|1,1;1,2;1,3;1,4|
|Temporal.Asynchronous.Succession|||||||
202..257|1,0,0;1,0,1,0;1,0,1,1;1,0,1,3|
Inh|Null|Null|Null|||263..282|1,0,1,2,1|
Inh|Null|Null|Null||||
```

The design methodology outlined in Ide and Bunt (2010) consists of a two-phase process: the specification of (1) an **abstract syntax** consisting of a *conceptual inventory* of the elements from which these structures are built up, and *annotation construction rules*, which describe the possible combinations of these elements into annotation structures; and (2) specification of at least one **concrete syntax** providing physical representations for these structures. This methodology has evolved in the context of developing standardized linguistic annotation schemes within ISO TC37 SC4, the foundation of which is the Linguistic Annotation Framework (LAF), (Ide and Romary, 2004); ISO 24612, 2009. LAF defines an *abstract model* for annotations consisting of a directed graph decorated with feature structures that is realized concretely in an XML serialization, the Graph Annotation Format (GrAF), (Ide and Suderman, 2007). GrAF serves as a *pivot* format into which well-formed annotation schemes may be mapped, thus guaranteeing syntactic consistency and completeness for the purposes of comparison, merging, and transduction to other formats.

In the context of ISO work, the abstract syntax for a given annotation type is developed before

[2] A format conversion tool is available from the PDTB Tools site: http://www.seas.upenn.edu/~pdtb/PDTBAPI/

any concrete realization is specified. However, because the PDTB annotation scheme already exists, we must "reverse engineer" the abstract syntax. The PDTB scheme is a flat structure in which the same information types are repeated multiple times in order to associate them with different annotation elements. For example, values for the same set of features are given for the connective, and the two arguments. In the conceptual inventory, a structure providing this information will be defined once and re-used where necessary.

An abstract syntax for the PDTB annotations includes the following conceptual inventory:

- A set of discourse relations REL = {*explicitRelation, implicitRelation, alternativeLexicalRelation, entityRelation, noRelation*};
- A pair of arguments ARGS = {ARG_1, ARG_2};
- A finite set F of attribution features = {*Source, Type, Polarity, Determinacy*} where *Source* ={Writer*, Other, Arbitrary, Inherited}; *Type* = {Comm (assertion)*, PAtt (belief), Ftv (fact), Ctrl (action), Null}; *Polarity* = {Null*, Negative}; *Determinacy* = {Null*, Indeterminate} (starred values are defaults).
- a semantic class CLASS;
- A connective head HEAD, consisting of the textual rendering of the head of a connective and its semantic class;
- Implicit connective IC, consisting of the textual rendering of the head of a connective and its semantic class;
- Attribution ATTR, consisting of a textual rendering of an explicit attribution, which may be empty;
- Supplementary text SUP, a reference to a span or spans in the text, which may be empty.

The annotation construction rules are the following:

- An **entityRelationEntity** is a pair $\langle a_1, a_2 \rangle \in$ ARGS.
- An **noRelationEntity** is a pair $\langle a_1, a_2 \rangle \in$ ARGS.
- An **explicitRelationEntity** is a triple $\langle EF, arg_1, arg_2 \rangle$, where EF is an explicitConnectiveStructure.
- An **implicitRelationEntity** is a triple $\langle IF, arg_1, arg_2 \rangle$, where IF is an implicitConnectiveStructure.
- An **altLexRelationEntity** is a triple $\langle AF, arg_1, arg_2 \rangle$, where AF is an altLexConnectiveStructure.

where arg_1, arg_2 are each an **argumentEntity Structure**; other structures defined as follows:

- An **explicitConnectiveStructure** is a pair ⟨*AttrF, HEAD*⟩, where HEAD is a pair ⟨*TEXT, CLASS*⟩;
- An **altLexConnectiveStructure** is a pair ⟨*AttrF, CLASS*⟩;
- an **implicitConnectiveStructure** is a pair ⟨*AttrF, IC*⟩ or a triple ⟨*AttrF, IC_1, IC_2*⟩, where IC is a pair ⟨*TEXT, CLASS*⟩;
- An **argumentEntityStructure** is a triple ⟨*a, AttrF,SUP*⟩ where a ∈ ARGS.

In all of the above, AttrF = (ATTR, F) where ATTR is a text span or empty, as defined above; and F = {*s,a,n,i*} with $s \in$ Source, $a \in$ Type, $n \in$ Polarity, $i \in$ Indeterminacy.

4 Concrete syntax

Based on the abstract syntax, a concrete syntax can be defined that provides a physical representation of a PDTB annotation. To ensure meaning-preserving mappings, there should be a one-to-one correspondence between the structures of the concrete syntax and those defined by the corresponding abstract syntax. Correspondingly, the concrete syntax should be mappable to a meaning-preserving representation in GrAF.

Figure 1 shows a concrete XML syntax for the PDTB annotation. Note that explicit and altLex entities, arguments, and supplemental information are anchored in the text by direct references to positions in the primary data (*WSJ* raw text files) and also by Gorn addresses that refer to node(sets) in Penn Treebank constituency trees. Implicit relations, entity relations, and noRels are associated with an *inferenceSite*, which give the character offset of the first character of arg_2 and its sentence number in the primary data.

This structure maps trivially to a GrAF rendering, given in Figure 7. The resulting graph is depicted in Figure 2. Note that the GrAF rendering requires that direct references to primary data in the XML annotation refer instead to *regions* defined in a *base segmentation* document. Such a

```
<explicitRelation
   span="258..262"
   gorn="1,0,1,2,0"
   src="Ot">
  <head sClass="Temp.Asynch.Succ">
      once</head>
  <arg1 span="202..257" gorn="1,0,0;
       1,0,1,0;1,0,1,1;1,0,1,3"
       src="Inh"/>
  <arg2 span="263..282"
        gorn="1,0,1,2,1" src="Inh"/>
</explicitRelation>
```

Figure 1: PDTB Concrete XML syntax

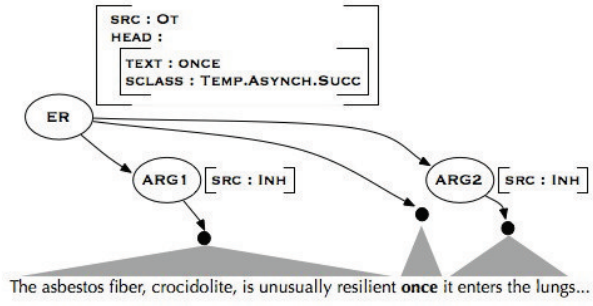


Figure 2: Graphic rendering of PDTB annotation

document would contain the minimal set of regions used by annotations over the data. So, for example, a GrAF rendering of the PTB syntactic annotations, PropBank semantic role annotations,[3] and PDTB annotations over the *WSJ* would refer to the same set of regions, combining several if necessary to refer to less granular or discontiguous spans. This avoids problems of alternative tokenizations, which in turn facilitates the combination of the different layers.

Figure 3 shows the concrete syntax realization of an implicit discourse relation for the text :

6. A Lorillard spokeswoman said, "This is an old story. *We're talking about years ago before anyone heard of asbestos having any questionable properties.* Implicit=besides **There is no asbestos in our products now**." (Expansion.Conjunction/Comparison)

Figure 4 shows the concrete syntax realization of an entity relation for the text :

7. *We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.* EntRel **Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University**.

[3]The ANC Project has developed PTB-to-GrAF and PropBank-to-GrAF transducers.

```
<implicitRelation
   src="Ot" type="Comm">
   <attr span="726..753"
     gorn="4,0;4,1,0;4,1,1;4,1,2;4,2"/>
  <ic sClass1="Expansion.Conjunction"
      sClass2="Comparison">
      besides</ic>
  <arg1 span="778..874" gorn="5"
       src="Inh"/>
  <arg2 span="876..916"
       gorn="6" src="Inh"/>
</implicitRelation>
```

Figure 3: Concrete syntax for an implicit discourse relation

```
<entityRelation>
<arg1 span="1046..1169"
        gorn="8"/>
  <arg2 span="1171..1311"
        gorn="9"/>
</entityRelation>
```

Figure 4: Concrete syntax for an entity discourse relation

Figures 5 and 6 show their graphic renderings.

5 Discussion

The exercise of creating an abstract syntax for the PDTB annotation scheme and rendering it in a graphic form shows the structure of the annotations clearly. The concrete syntax is much more readable than the original format, and therefore errors and inconsistencies may be more readily identified. Furthermore, because it is rendered in XML, annotations can be validated against an XML schema (including validation that attribute values are among a list of allowable alternatives).

The abstract syntax also suggests an overall structure for a general-purpose standard for annotating discourse relations, in that it identifies a

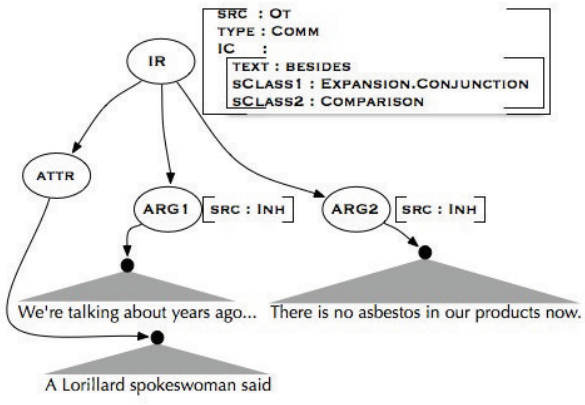


Figure 5: Implicit relation visualization

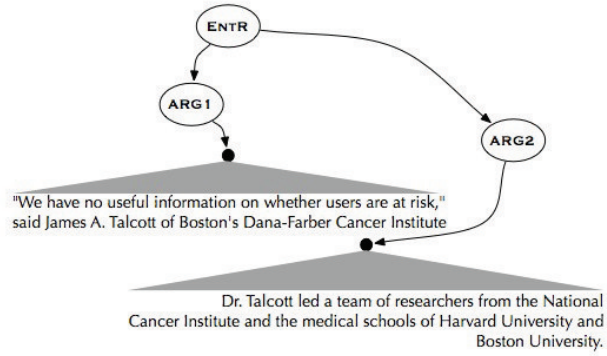


Figure 6: Entity relation visualization

high-level set of relations based on textual realization (the set REL, as given in Section 3) that could provide the top-level of a relation classification scheme. We envision that any general-purpose discourse annotation scheme must allow for annotation based on all or any of several perspectives on elements of the task, such as semantic, interpersonal/intentional, and stylistic/textual identified in Hovy (1995). PDTB annotations are classified as "informational" (semantic, inter-propositional, ideational, pragmatic); the intentional and textual perspectives lie outside the scope of PDTB. PDTB's attribution types and the set of semantic classes[4], combined with those of other schemes, could provide a base for development of a structured set of discourse annotation classes for the ISO specification along the various axes of perspective, and at different levels of granularity.

The work within ISO on discourse relations so far focuses on *discourse graphs* and *discourse trees* that describe discourse structure over an entire text by linking individual relations. Annotating dependencies across relations, however, presumes an understanding of the nature of representation for high-level discourse structure, so that the resulting theory can guide the specification of guidelines for how to compose individual relations. Since there is currently little agreement on such a theory, the PDTB has taken the approach to avoid biasing the annotation towards one or the other theory. Instead, the developers have chosen to specify discourse relations at a low-level, i.e., a level that is clearly defined and well understood. This aspect of the methodology has two related benefits. First, the corpus is usable by researchers across different frameworks for empirical studies on the behavior of low-level discourse relations (for example, studies on the differences between causal relations and contrastive relations). Second, the underlying theory-neutrality itself will allow validation studies of different types of discourse representations (e.g., trees (Polanyi, 1987; Mann and Thompson, 1988; Webber et al., 2003), graphs (Wolf and Gibson, 2005), DAGs (Lee et al., 2008)). In this sense, the PDTB uniquely provides a basis for an *emergent* and *data-driven* theory of discourse structure. Consideration of this approach, either as an alternative to full discourse trees/graphs, or as a base level upon which to build higher-level representations of various types, could be valuable input for the ISO development of a general-purpose standard.

6 Conclusion and Future Work

We have developed an interoperable format for the Penn Discourse Treebank annotation scheme using the strategy for scheme design outlined in Ide and Bunt (2010). We show a high-level XML concrete syntactic realization of the abstract syntax and the corresponding GrAF representation, which ensures that PDTB annotations can be easily combined with other GrAF-based annotation layers in the same source corpus. We are developing a transducer to render the current PDTB annotations in the XML representation, which can be transduced to GrAF using the ANC Tool[5]. This transducer will be made freely available on the PDTB web site.

The underlying annotation framework of the PDTB is followed by several similar discourse annotation projects, in that they are based on a similar discourse annotation framework (e.g., lexicalization of discourse relations, structural independence of relations, and theory neutrality). However, among these schemes there are variations (e.g., the inventory of sense classes, feature sets for attribution, and relation types) dictated by the language and/or domain of the data being annotated. By abstracting out the basic elements of the PDTB scheme and formalizing the structure of the information in a PDTB annotation, we have identified some of the conceptual building blocks of discourse relation analysis that can be used to guide development of new schemes as well as compari-

[4]For a complete list, see the "Penn Discourse Treebank 2.0. Annotation Manual", http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

[5]http://www.anc.org/tools

```
<node xml:id="pdtb-n101"/>
<a label="explicitRelationEntity"
   ref="pdtb-n101" as="PDTB">
<fs>
  <f name="src" value="Ot"/>
</fs>
</a>

<node xml:id="pdtb-n12"/>
<a label="explicitRelationEntity"
   ref="pdtb-n11" as="PDTB">
<fs>
  <f name="src" value="Ot"/>
  <f name="head">
     <fs>
        <f name="text" value="once"/>
        <f name=sClass"
            value="Temp.Asynch.Succ"/>
     </fs>
   </f>
</fs>
</a>

<edge xml:id="pdtb-e201" from="pdtb-n12"
   to="seg-r15 seg-r16"/>
<edge xml:id="pdtb-e202" from="pdtb-n12"
   to="pdtb-n13"/>
<edge xml:id="pdtb-e203" from="pdtb-n12"
   to="pdtb-n14"/>
<edge xml:id="pdtb-e204" from="pdtb-n12"
   to="pdtb-n15"/>


<node xml:id="pdtb-n14"/>
<a label="arg1" ref="pdtb-n14"
  as="PDTB">
<fs>
  <f name="src" value="Inh"/>
</fs>
</a>

<edge xml:id="pdtb-e205" from="pdtb-n14"
   to="seg-r20"/>

<node xml:id="pdtb-n15"/>
<a label="arg2" ref="pdtb-n15"
  as="PDTB">
<fs>
  <f name="src" value="Inh"/>
</fs>
</a>

<edge xml:id="pdtb-e206" from="pdtb-n15"
   to="seg-r57 seg-r65"/>
```

Figure 7: GrAF rendering of PDTB example

son and combination of schemes. Our immediate goal is to explore the potential of this work to feed the development of an ISO standard for annotating discourse relations.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90, Morristown, NJ, USA. Association for Computational Linguistics.

Harry Bunt. 2010. A methodology for designing semantic annotation languages exploiting semantic-syntactic isomorphisms. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL2010)*, pages 29–46, Hong Kong SAR. City University of Hong Kong.

Saul Gorn. 1965. Explicit definitions and linguistic dominoes. In John Hart and Satoru Takasu, editors, *Systems and Computer Science*. University of Toronto Press, Toronto, Canada.

Eduard Hovy. 1995. The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers*, The Netherlands. Egmond-aan-Zee.

Nancy Ide and Harry Bunt. 2010. Anatomy of annotation schemes: Mapping to GrAF. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 247–255, Uppsala, Sweden, July. Association for Computational Linguistics.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Journal of Natural Language Engineering*, 10(3–4):211–225.

Nancy Ide and Keith Suderman. 2007. GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the First Linguistic Annotation Workshop*, pages 1–8, Prague.

ISO. 2009. *Semantic annotation framework (SemAF), Part 1: Time and Events. ISO CD 24617-1:2009*. ISO, Geneva.

Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2008. Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse III Workshop*, Potsdam, Germany.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Livia Polanyi. 1987. The linguistic discourse model: Towards a formal theory of discourse structure. Technical Report 6409, Bolt Beranek and Newman, Inc., Cambridge, Mass.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

PDTB-Group. 2008. The Penn Discourse TreeBank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2).

FIELD NUM	DESCRIPTION
1.	*Relation Type:* Encodes how the relation is realized in the text: {Explicit, Implicit, AltLex, EntRel, NoRel}
2.	*Conn Span:* Character offsets (or set of offsets for discontinuous text) for the connective or the AltLex text span
3.	*Conn Head:* Head of Explicit connective (provided as text)
4.	*Conn Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Conn Span"
5.	*Conn Attr Src:* Attribution Source features for Connective: {Wr (Writer), Ot (Other), Arb (Arbitrary)}
6.	*Conn Attr Type:* Attribution Type features for Connective: {Comm (assertion), PAtt (belief), Ftv (Fact), Ctrl (action)}
7.	*Conn Attr Pol:* Attribution Polarity feature for Connective: {Neg (negative polarity), Null (no negation)}
8.	*Conn Attr Det:* Attribution Determinacy feature for Connective: {Indet (indeterminate), Null (determinate)}
9.	*Conn Attr Span:* Character offsets (or set of offsets for discontinuous text) for text span of attribution on connective
10.	*Conn Attr Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Conn Feat Span"
11.	*Conn1:* First implicit connective
12.	*SClass1A:* First sense for explicit connective or (first) implicit connective
13.	*SClass1B:* Second sense for explicit connective or (first) implicit connective
14.	*Conn2:* Second implicit connective
15.	*SClass2A:* First sense for second implicit connective
16.	*SClass2B:* Second sense for second implicit connective
17.	*Sup1 Span:* Character offsets (or set of offsets for discontinuous text) for supplemantary text for Arg1
18.	*Sup1 Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Sup1 Span"
19.	*Arg1 Span:* Character offsets (or set of offsets for discontinuous text) for Arg1 text span
20.	*Arg1 Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Arg1 Span"
21.	*Arg1 Attr Src:* Attribution Source features for Arg1: {Wr (Writer), Ot (Other), Arb (Arbitrary), Inh (inherited)}
22.	*Arg1 Attr Type:* Attribution Type features for Arg1: {Comm (assertion), PAtt (belief), Ftv (Fact), Ctrl (actions), Null}
23.	*Arg1 Attr Pol:* Attribution Polarity feature for Arg1: {Neg (negative polarity), Null (no negation)}
24.	*Arg1 Attr Det:* Attribution Determinacy feature for Arg1: {Indet (indeterminate), Null (determinate)}
25.	*Arg1 Attr Span:* Character offsets (or set of offsets) for text span of attribution on Arg1
26.	*Arg1 Attr Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Arg1 Feat Span"
27.	*Arg2 Span:* Character offsets (or set of offsets for discontinuous text) for Arg2 text span
28.	*Arg2 Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Arg2 Span"
29.	*Arg2 Attr Src:* Attribution Source features for Arg2: {Wr (Writer), Ot (Other), Arb (Arbitrary), Inh (inherited)}
30.	*Arg2 Attr Type:* Attribution Type features for Arg2: {Comm (assertion), PAtt (belief), Ftv (Fact), Ctrl (actions), Null}
31.	*Arg2 Attr Pol:* Attribution Polarity feature for Arg2: {Neg (negative polarity), Null (no negation)}
32.	*Arg2 Attr Det:* Attribution Determinacy feature for Arg2: {Indet (indeterminate), Null (determinate)}
33.	*Arg2 Attr Span:* Character offsets (or set of offsets) for text span of attribution on Arg2
34.	*Arg2 Attr Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Arg2 Feat Span"
35.	*Sup2 Span:* Character offsets (or set of offsets for discontinuous text) for supplementary text for Arg2
36.	*Sup2 Gorn:* PTB Gorn address (or set of addresses when text is not covered by a single node) for "Sup2 Span"

Table 1: Annotation Fields for the Penn Discourse Treebank Flat (Horizontal) Format

Author Index